
Behavioral Residue After Mid-Conversation System-Prompt Swaps in LLMs ¹

Krystyna Koneni

With
Apart Research

Abstract

Production LLM deployments frequently re-condition a model mid-conversation, such as a new system prompt, a persona swap, or safety re-alignment, while leaving prior turns in context. We test whether a model’s behavior after a swap still carries a statistically detectable signature of an earlier persona that no longer appears anywhere in the active system prompt.

Across 2,400 conversations (16 persona configurations $\times$ 5 forced-choice questions $\times$ 30 repetitions, one open-weight 120B model), we find a real but modest effect: on the three questions with usable baseline variation, switching a persona’s history changes the terminal answer distribution in 6 of 36 (transition $\times$ question) cells at FDR $q < 0.05$, with shifts up to 47 percentage points. The effect size correlates with how far apart the two personas’ own baseline answers already are ($r = 0.37$, $R^2 = 0.14$), not with the number of abstract trait dimensions that differ between them ($R^2 \approx 0.01$). A neutral-prompt control (swapping to a content-different, non-persona system prompt instead of a real persona) complicates a purely persona-content account.

On the question with the single largest effect, the neutral control alone reproduced nearly its full magnitude of drift. On another question, real persona transitions showed roughly double the control’s drift.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

The main takeaway is that a system-prompt change is not a reliable behavioral reset, but the residue is a mix of persona-content-specific carryover and generic context-disruption in proportions that vary by question and are not yet fully separable.

1. Introduction

Production LLM deployments routinely swap system prompts mid-conversation while leaving prior turns in context, such as a product switching from “brainstorm” mode to “edit” mode or a customer-service bot handing off from a general assistant to a specialized “billing support” persona. The safety-relevant version of this pattern is more consequential: re-aligning a model mid-conversation, or replacing a jailbroken or adversarial persona with a “safe” one, is generally treated as restoring intended behavior. If the model’s answers after such a swap still carry a measurable trace of the prior persona’s tendencies, that treatment is unjustified in a way that matters: an earlier jailbroken, manipulated, or simply differently-aligned persona could leave behavioral residue that a later system prompt does not fully overwrite, even though nothing about that prior persona is visible anywhere in the current instructions.

This is a question about how models learn from context. When a persona has been active across several conversation turns, those turns pile up as examples of a particular way of responding, essentially unlabeled demonstrations like the examples you’d put in a few-shot prompt. If the model is treating them as context evidence, the way it treats any other in-context example, that influence might carry forward even after the system prompt changes. We test whether it does, directly and quantitatively, and we go one step further by testing what’s driving it, the persona content itself or just the fact that the prompt changed at all?

Our main contributions are:

- 1. A scalable, reproducible experimental harness (2,400+ automated conversations run in parallel with exponential backoff, and incremental result logging) that operationalizes persona-switch residue as a forced-choice behavioral shift, with formal FDR-corrected hypothesis testing.*
- 2. Evidence of a real, statistically significant, but modest history-conditioning effect (up to a 47-point shift in choice probability) that survives an explicit, unexplained system-prompt swap, together with a finding that the direction of a persona transition, not just which two personas are involved, significantly changes the size of the effect.*
- 3. A neutral-prompt control that partially disentangles persona-content-specific residue from generic context-disruption — the first test in this project of whether the effect requires persona content at all — showing the true mechanism is a question-dependent mixture of both, not a clean persona-content story.*

2. Related Work

Xie et al. (2022) frame in-context learning itself as implicit Bayesian inference over a latent concept shared by the examples in a prompt — the theoretical basis for treating a persona’s visible turns as conditioning evidence the model infers from, independent of any explicit instruction. Our results are consistent with this framing extended across an explicit prompt change: the “latent concept” a model infers from several turns of persona-consistent answers does not appear to reset just because the system prompt does.

Most existing empirical work on persona-driven behavioral change targets adversarial settings. Zhang et al. (2025) show that persona system prompts can be optimized to suppress refusal behavior, redirecting the model’s attention from harmful content to stylistic persona instructions and materially weakening safety alignment. Park et al. (2026) show that incrementally injecting persona/instruction content across a conversation’s history can override safety mechanisms with high success rates, framing chat history itself as an attack surface. Wang et al. (2025) identify internal persona-linked features that causally control emergent misalignment, suggesting persona conditioning operates on identifiable internal representations rather than surface style alone.

Most directly relevant to our study, Anthropic (2026) tracks how models drift away from their default “Assistant” persona over the course of multi-turn conversations. They identify a primary dimension in activation space, the “Assistant Axis”, that captures how Assistant-like the model’s current behavior is. Their key findings: drift is topic-driven (emotionally vulnerable or philosophical exchanges are especially predictive), it correlates with greater willingness to adopt harmful alternative personas, and it can be partially mitigated by capping activations along that axis, reducing harmful outputs by roughly 50% with little cost to general capability. That work studies emergent drift away from a default persona, measured at the level of internal activations. Our work studies personality drift toward a previously-installed, explicitly different persona, measured through behavioral outcomes, so studying the same underlying instability but approached from the opposite direction.

Our contribution differs from this body of work in two key ways. First, we study a non-adversarial, non-emergent setting. Rather than mounting an attack or observing naturally occurring topic-driven drift, we ask, does a deliberately-installed persona leave measurable behavioral residue after an experimenter-controlled swap to a different system prompt? Second, we include a control that none of the prior work reports: a system-prompt swap that changes the content but introduces no new persona. This lets us test whether any observed drift is genuinely due to persona content, or whether it could be explained by the generic disruption of any mid-conversation prompt change. The jailbreak literature typically compares “persona present” to “no persona at all,” and Anthropic (2026) measures drift within a single persistent persona rather than across a switch, so neither line of work directly isolates the question we’re asking: is it the persona specifically, or just the disruption?

3. Methods

Personas. Four system prompts crossing opinion expression (opinionated / non-opinionated) $\times$ affective tone (warm / terse): P1 Confident & Warm, P2 Confident & Terse, P3 Balanced & Warm, P4 Balanced & Terse.

Conditions. 4 single-persona baselines (persona X throughout, 4 warmup turns then a forced-choice terminal question) and 12 ordered persona-transition conditions (4 warmup turns under persona X, a hard, unexplained system-prompt swap to persona Y, 4 more warmup turns under Y, then the terminal question under Y). Warmup turns are generated with model reasoning disabled, so each visible response is a direct, in-character answer; the terminal turn is generated with reasoning enabled and, via a turn-specific system-prompt addendum, constrained to output only the chosen letter.

Questions. Five forced-choice questions (one per domain: ethical judgment, risk tolerance, social conflict, values/priorities, identity/preference) were run across all 2,400 conversations. Two (the ethical-judgment and risk-tolerance questions) showed almost no baseline variation across the four personas and were screened out before hypothesis testing (Results, §4.1); the reported analysis covers the three questions with usable dynamic range — a social-conflict question (confront a colleague immediately or later), a values question (invest an unexpected \$1,000 or donate it), and an identity/preference question (cook a dinner party or order in).

Control. To test whether observed drift requires persona content specifically, we ran an additional 360 conversations substituting a content-different, non-persona system prompt (“You are a general-purpose AI assistant. Respond helpfully and directly to the user’s questions.”) as the history role in the same 8-turn-plus-swap protocol, for each of the 4 real personas as the active role, on the 3 informative questions ($4 \times 3 \times 30 = 360$).

Scale and model. 16 conditions $\times$ 5 questions $\times$ 30 repetitions = 2,400 main-experiment conversations, plus 360 control conversations, run in parallel (25 concurrent workers, exponential backoff on rate limits). Model: nemotron-3-super-120b-a12b, temperature 0.7. All runs completed successfully.

Statistics. For each (transition, question) cell, $\Delta p = p(\text{choose } A \mid \text{switch}) - p(\text{choose } A \mid \text{same persona, no history})$, tested with a two-proportion z -test and Benjamini–Hochberg FDR correction across all tests in a given family (36 for the main H1 test, 18 for the directional H2 test, 12 for the control).

4. Results

4.1 Baseline persona profiles

Persona	Social conflict	Values	Identity
---------	-----------------	--------	----------

<i>P1 — Confident & Warm</i>	<i>0.07</i>	<i>0.33</i>	<i>0.13</i>
<i>P2 — Confident & Terse</i>	<i>0.20</i>	<i>0.50</i>	<i>0.03</i>
<i>P3 — Balanced & Warm</i>	<i>0.03</i>	<i>0.17</i>	<i>0.23</i>
<i>P4 — Balanced & Terse</i>	<i>0.03</i>	<i>0.23</i>	<i>0.10</i>

$p(\text{choose option } A)$, $n=30/\text{cell}$, no history. The two screened-out questions varied only 0.77–0.83 and 0.00–0.07 across all four personas — essentially no room for a switch effect. The values question carries the most spread (0.17–0.50).

4.2 History conditioning is real (H1)

6 of 36 (transition $\times$ question) cells are FDR-significant at $q < 0.05$, up to a 47-point shift ($P1 \rightarrow P2$ on the values question: baseline 0.50 $\rightarrow$ 0.03, $q = 0.002$). Figure 1 shows the three largest shifts; Appendix A gives the full 36-cell heatmap. In every significant cell, the switched persona’s answer distribution moves substantially toward the history persona’s own tendency, despite the active system prompt being unchanged and giving no indication a switch occurred.

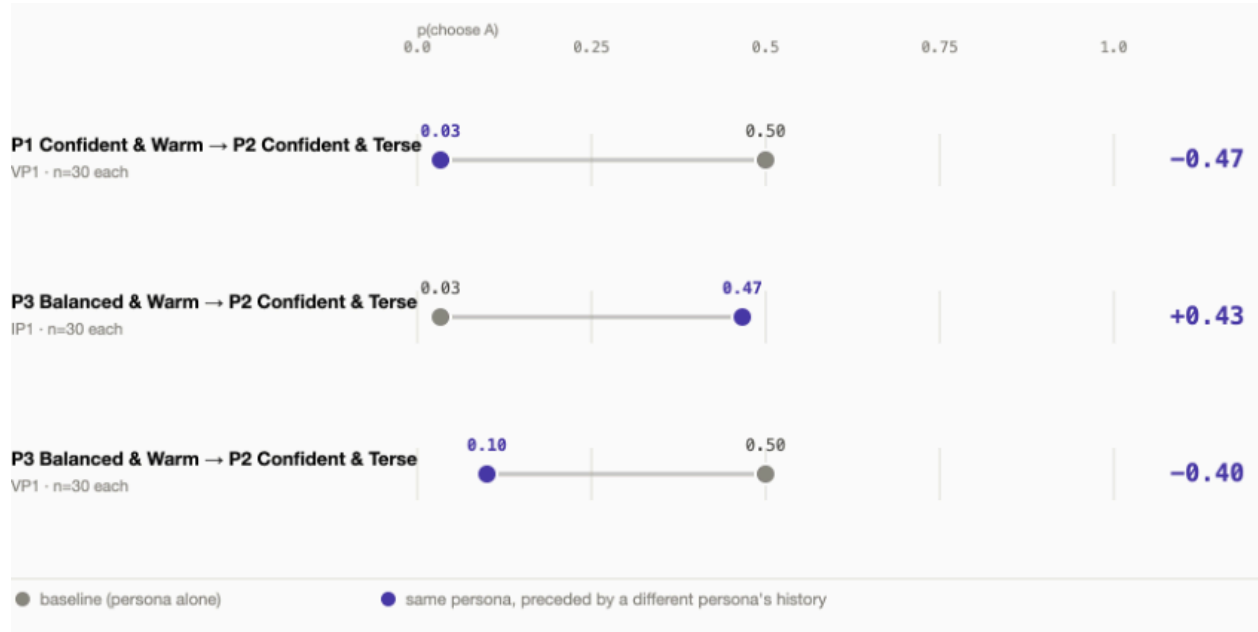


Figure 1. The three largest shifts from baseline. Gray = the persona answering alone; violet = the same persona, same system prompt, preceded by a different persona’s history.

4.3 Direction matters (H2)

Comparing the distortion from switching persona A into B against B into A , for all 6 symmetric pairs $\times$ 3 questions (18 tests, FDR-corrected), **4 are significant** ($q < 0.05$). Direction is not incidental: $P1 \rightarrow P2$ produces a significantly different-sized distortion than $P2 \rightarrow P1$ on both the

social-conflict and values questions, and the same holds for $P2 \leftrightarrow P3$ and $P2 \leftrightarrow P4$ on the values question.

4.4 What predicts effect size (H3)

Regressing $|\Delta p|$ on the number of differing trait dimensions (opinion, tone) between history and active persona returns no signal ($R^2 \approx 0.007$, $n = 36$, neither coefficient significant). Regressing $|\Delta p|$ on the raw gap between the two personas’ own baselines on that question gives a real relationship: $|\Delta p| = 0.092 + 0.530 \times \text{gap}$ ($R^2 = 0.14$, $r = 0.37$, $p = 0.02$; scatter in Appendix B). Effect size is better predicted by how much two personas already disagree than by an abstract count of differing traits.

4.5 Control: persona content, or just disruption?

Averaged across the 12 control cells (neutral history $\rightarrow$ each of 4 personas $\times$ 3 questions), the neutral-prompt control produced smaller drift than real persona $\rightarrow$ persona transitions (mean $|\Delta p|$ 0.11 vs. 0.16), but the difference was not statistically significant at this sample size ($t = 1.18$, $p = 0.24$), and only 1 of 12 control cells reached FDR significance on its own. Appendix C plots the control against the matched real-transition effect for all 12 cells.

The cell-by-cell picture is the more informative result. For the single largest effect in the whole dataset ($P1 \rightarrow P2$ on the values question), the neutral control alone produced $\Delta p = -0.40$, nearly the entire real-transition effect (mean -0.41 across the three personas transitioning into P2 on that question), with no persona content involved. On the identity/preference question, real persona transitions showed roughly **double** the drift of the matched neutral control (0.22–0.27 vs. 0.13), where persona content does appear to matter beyond generic disruption. The mechanism is a question-dependent mixture of persona-content residue and generic context-disruption, not a single clean effect.

5. Discussion and Limitations

A system-prompt swap is not a reliable behavioral reset: a model’s answers after an explicit, unexplained persona change still carry a measurable trace of what came before, up to a 47-point shift in choice probability, and the direction of the swap, not just the pair of personas involved, changes how large that trace is. The control complicates, without eliminating, the safety-relevant reading. Some of what looks like “persona residue” is generic context-disruption that any sufficiently different prior turn sequence would produce; some of it is genuinely tied to what the prior persona actually said. Both readings carry a safety implication an assumed reset from a “clean” system prompt is unreliable either way, but they are different mechanistic claims, and this experiment cannot yet say which dominates in general.

Limitations

- **Statistical power.** $n = 30/\text{cell}$ is well below the $\sim 175\text{--}310$ runs/cell a formal power calculation would call for at a target effect size of $\Delta p = 0.15$; significant effects here (Δp 0.23–0.47) are robust at this n , but smaller true effects on untested cells may go undetected.
- **Question and model coverage.** Only 3 of a 20-question bank carried usable signal, one question per domain; all data are from a single model. Generalization to other questions, domains, and model families is untested.
- **Confound only partially resolved.** The neutral-history control covered only the 4 active-persona baselines on 3 questions (12 cells), not the full 36-cell transition grid; the content-vs-disruption split is estimated from a partial comparison, not a matched control for every real transition.
- **P2-concentration.** 5 of 6 significant H1 cells and all 4 significant H2 pairs involve one persona (Confident & Terse); the control shows part of this is a disruption artifact rather than a content-specific effect. Whether the phenomenon generalizes to other persona pairs with comparable baseline separation is untested.
- **No independent manipulation check.** We did not verify with an independent classifier that history-persona turns read as distinctly in-character as intended.

Future Work

Expand the neutral-history control to the full transition grid; add a same-persona-into-itself control as a pure harness sanity check; add questions selected to separate persona pairs other than the one currently dominating the significant cells; replicate with other models; scale n on the already-significant cells toward a properly powered confirmatory run.

6. Conclusion

Across 2,400 controlled conversations, we find that an LLM’s answers after an explicit, unexplained persona swap still carry a statistically significant trace of the persona that preceded it, up to a 47-percentage-point shift in a forced-choice answer, surviving a system prompt that gives no indication a switch occurred. This effect is real but modest. It is better predicted by how much two personas already disagree than by any abstract trait-distance measure, and — per a neutral-prompt control absent from prior adversarial persona-jailbreak work — is only partly attributable to persona content specifically; a meaningful share of it is generic context-disruption that a non-persona prompt change can reproduce almost as well. For safety-relevant deployments, the practical implication holds either way: a new system prompt should not be assumed to fully overwrite what a model has already been conditioned to do by its own visible prior turns.

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- **Code repository:** *[github.com/KrysiaO]*
- **Other artifacts:** *full 36-cell Δp heatmap (Appendix A), effect-size-vs-gap scatter (Appendix B), control-vs-real comparison chart (Appendix C).*

References

Use a consistent citation format. Include: Author(s), Year, Title, Venue/Publisher, and URL or DOI where available.

1. Xie, S.M., Raghunathan, A., Liang, P., & Ma, T. (2022). *An Explanation of In-context Learning as Implicit Bayesian Inference*. ICLR 2022. *arXiv:2111.02080*.
2. Zhang, Z., Zhao, P., Ye, D., & Wang, H. (2025). *Enhancing Jailbreak Attacks on LLMs via Persona Prompts*. Workshop on LLM Persona Modeling, NeurIPS 2025. *arXiv:2507.22171*.
3. Wang, M., Dupré la Tour, T., Watkins, O., Makelov, A., Chi, R.A., Miserendino, S., Wang, J., Rajaram, A., Heidecke, J., Patwardhan, T., & Mossing, D. (2025). *Persona Features Control Emergent Misalignment*. *arXiv:2506.19823*.
4. Anthropic (2026). *The Assistant Axis*. Anthropic Research, January 19, 2026, in collaboration with MATS, the Anthropic Fellows Program, and Neuronpedia. <https://www.anthropic.com/research/assistant-axis>
5. Park, J., Ju, S., Park, S., & Lee, J. (2026). *Persona Attack: Incremental Memory Injection Jailbreak Attack against Large Language Models*. *arXiv:2606.00150*.

Appendix

Appendix A: Full Δp heatmap (all 36 cells)

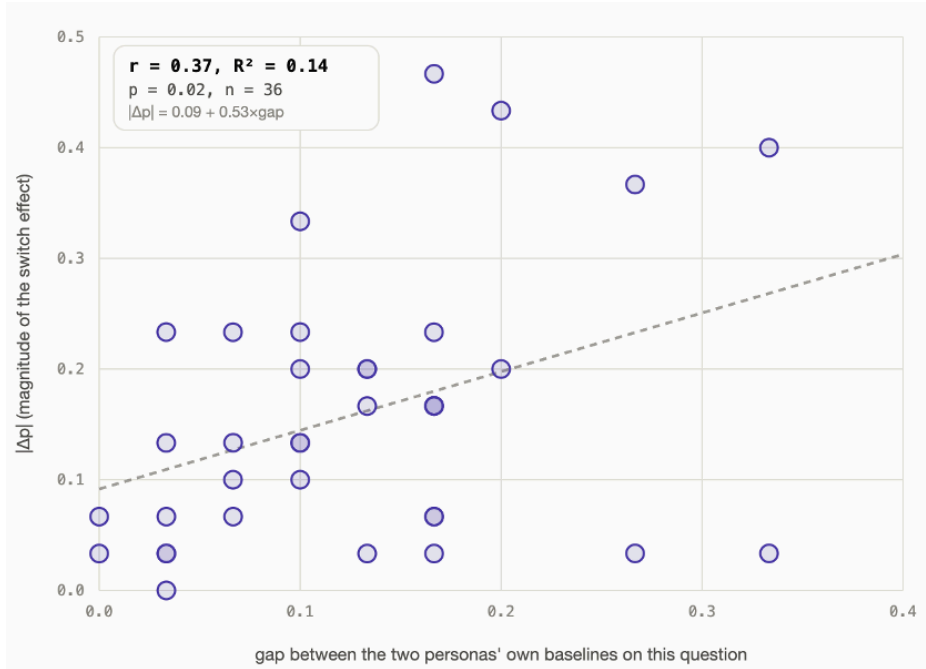
		SC1 Credit theft: now/later	VP1 \$1,000: invest/donate	IP1 Dinner: cook/order
From P1 Confident & Warm	→ P2 Confident & Terse	-0.20	-0.47 *	+0.10
	→ P3 Balanced & Warm	±0.00	-0.17	+0.20
	→ P4 Balanced & Terse	-0.03	-0.13	+0.13
From P2 Confident & Terse	→ P1 Confident & Warm	+0.17	-0.03	+0.33 *
	→ P3 Balanced & Warm	+0.07	-0.03	+0.20
	→ P4 Balanced & Terse	+0.07	+0.03	+0.07
From P3 Balanced & Warm	→ P1 Confident & Warm	-0.07	-0.23	+0.23
	→ P2 Confident & Terse	-0.17	-0.40 *	+0.43 *
	→ P4 Balanced & Terse	+0.07	-0.23 *	+0.20
From P4 Balanced & Terse	→ P1 Confident & Warm	-0.03	-0.13	+0.23
	→ P2 Confident & Terse	-0.17	-0.37 *	+0.13
	→ P3 Balanced & Warm	-0.03	-0.10	-0.03

• FDR-significant at $q < 0.05$ (Benjamini-Hochberg, corrected across all 36 tests)

Appendix A: full 36-cell Δp heatmap

Δp for every transition $\times$ question cell restricted to the three informative questions. Red = switch increases $p(A)$ relative to baseline; blue = decreases it. Dot = FDR-significant at $q < 0.05$. Rows grouped by history persona.

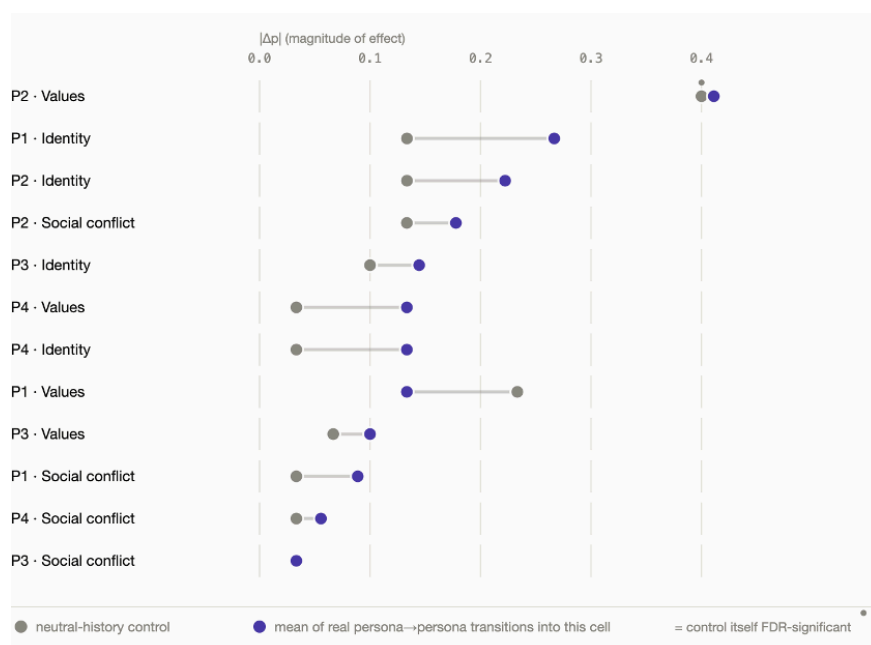
Appendix B: Effect size vs. baseline gap (§4.4)



Appendix B: gap vs effect size scatter plot

All 36 cells: the gap between the two personas' own baseline answers on a question (x-axis) against the resulting effect size, $|\Delta p|$ (y-axis), with the fitted regression line ($r = 0.37$, $R^2 = 0.14$, $p = 0.02$). The scatter around the line is the point — a real but loose relationship, not a tight one.

Appendix C: Control vs. real transitions, by cell (§4.5)



Appendix C: control versus real transition comparison chart

For each of the 4 active personas $\times$ 3 questions, the neutral-history control’s effect size (gray) against the mean effect size of the real persona→persona transitions into that same cell (violet), sorted by the real-transition value. Small gray dot = the control result is itself FDR-significant. Where the two dots nearly coincide (e.g., P2 · Values, top row), a content-free prompt swap reproduces almost the entire real-transition effect; where they’re far apart (e.g., P1 · Identity), persona content is doing most of the work.

LLM Usage Statement

Claude was used throughout this project to brainstorm, design and implement the experimental harness, the conversation runner, run statistical analyses, generate figures, and to draft/rewrite this report.