Coherence Without Contact

Fluent AI, false certainty, and the design of agency-preserving assistants.

PRE-REGISTERED STUDY DESIGN DRAFT V0.3 — NOT YET RESULTS JUNE 2026

ABSTRACT

Large language models can produce highly fluent, structurally coherent accounts regardless of whether those accounts are grounded. This paper argues that fluent generation can decouple two properties human readers often treat as linked: *felt coherence*, the sense that an account hangs together, and *contact*, the degree to which it is constrained by evidence, accuracy, uncertainty, and consequence. When these properties separate, confidence may track fluency rather than warrant, producing *false certainty*: subjective confidence unmatched by accuracy or support. We propose and test a design response, *completion-oriented, agency-preserving* assistants, which move users toward bounded, user-owned next steps rather than indefinite engagement or directive closure. Study 1 is designed to test the diagnostic dissociation using real LLM generations, controlled manipulations of felt coherence and contact, and a perturbation phase, extending recent calibration-gap findings. Study 2 is designed to test whether completion-oriented, agency-preserving interaction improves *durable agency*, operationalised as *48-hour action integrity*: whether users carry a self-specified, goal-aligned action, or a documented deliberate non-action, beyond the session.

KEYWORDS fluent AI; false certainty; calibration; metacognition; processing fluency; agency-preserving design; completion orientation; durable agency

CITATION Kok, C. (2026). *Coherence Without Contact: Fluent AI, False Certainty, and the Design of Agency-Preserving Assistants*. Draft v0.3 (pre-registered study design), with AI-assisted drafting.

Where this sits

This is a study design, not findings: its claims are framed to be tested and, where they fail, reported as disconfirmation. It is the human-science layer of a single idea the site approaches from three sides. The AI & Design doctrine states the design philosophy — build systems that help a user become clearer to themselves, and leave; this paper turns that into pre-registered, falsifiable studies. The Swarm Instrument measures the same coherence-vs-contact distinction inside a multi-agent system; this paper measures it inside the human reader. And the cut itself — coherence is not, by itself, contact — is the empirical face of the framework's treatment of a resolution as a conditioned coherence that is real but not, on that account, true.

1. Introduction

A user brings an unresolved question to an AI assistant. The answer comes back structured, fluent, and confident. None of that fluency is evidence that the answer is correct, because fluency is a property of the generation process, not of the world the answer is about.

In many everyday communicative contexts, coherence has functioned as a fallible but useful cue of contact, because producing a detailed, coherent account often required some knowledge of the subject. Fluent generation weakens that cue by making fluent, detailed accounts cheap to produce at arbitrary grounding, including none. We do not claim that consequential-but-ungrounded coherence is new; rhetoric, salesmanship, and propaganda are old. We claim that the cost, speed, and individualised, conversational delivery of fluent ungrounded accounts have changed, and that this matters for how human confidence calibrates. When certainty tracks the fluency of an account rather than its grounding, increases in fluency can produce increases in confidence independent of whether the account is correct.

This paper does two things. First, it is designed to diagnose that dissociation and to show it operates in an LLM-generated setting rather than only in researcher-written prose. Second, it asks whether interaction design can mitigate it. The dominant optimisation targets for deployed assistants are satisfaction and engagement, and both are exactly what a fluent-but-ungrounded answer is positioned to win. We propose that a different outcome, *durable agency* measured after the interaction ends, separates assistants that help people think from assistants that help people feel resolved, and that a specific design posture, completion-oriented and agency-preserving, improves it.

We state a commitment about the design claim in advance, in operational terms. The design hypothesis is the study's primary test, and the analysis plan specifies that a null or negative result on it is reported as the lead finding rather than as a caveat, with no post-hoc moderator analysis promoted above the pre-registered primary test. The details of that rule are given in Section 8. We make the commitment here so that the design claim can be disconfirmed cleanly rather than rescued.

2. Background

Four literatures bear on this, and part of the contribution is to connect them around generative AI and add a falsifiable design layer.

Processing fluency. People misread the *ease* of processing a stimulus as a signal about its *content*. Reber and Schwarz (1999) report that more perceptually fluent statements are judged true more often, and Alter and Oppenheimer (2009) frame fluency as a broad metacognitive cue affecting judgments of truth, familiarity, and liking. (Reber, Schwarz, and Winkielman (2004) establish fluency's relation to aesthetic pleasure and liking specifically, and we use it only for that narrower claim.) Fluency is a cue attributed to the wrong source, and it operates below deliberate judgment.

Automation bias and AI-assisted reliance. People over-accept automated outputs and under-engage their own scrutiny (Parasuraman and Manzey 2010). In AI-assisted decision-making, adding explanations can increase acceptance of AI recommendations regardless of correctness (Bansal et al. 2021). Two recent LLM-specific studies sharpen this. Kim et al. (2025) find that the presence of explanations increases reliance on both correct and incorrect LLM responses, while providing sources, or surfacing inconsistencies in explanations, reduces reliance on incorrect ones. Spatharioti et al. (2025) find that LLM-based search produces faster and more satisfying interactions with accuracy comparable to traditional search when the model is correct, but overreliance on incorrect output when the model errs, and that confidence-based highlighting helps users detect errors and improves decision accuracy. Most relevant to our design claim, Buçinca, Malaya, and Gajos (2021) show that *cognitive forcing functions*, designs that interrupt the path of least resistance and require the user to engage before accepting the AI's answer, reduce overreliance in AI-assisted decision support, at a measurable cost to subjective satisfaction. (Bansal et al. and Buçinca et al. concern AI decision support generally rather than LLMs specifically; we treat them as such.)

Calibration and the LLM calibration gap. Steyvers et al. (2025) establish the diagnostic neighbourhood directly. They define a *calibration gap*, the difference between human confidence in LLM answers and the models' actual confidence, and a *discrimination gap*, how well humans can distinguish the model's correct from incorrect answers. Users overestimated LLM accuracy when given default explanations, and longer explanations raised confidence without improving accuracy or discrimination. That last finding is felt coherence inflating certainty with no gain in contact, already measured. Study 1 is designed to extend this from their multiple-choice and short-answer setting into one with controlled felt-coherence and contact manipulations and an explicit perturbation phase.

Belief perseverance. For our perturbation prediction we draw on the finding that beliefs, once formed and given a coherent rationale, can persist even after their evidential basis is discredited (Ross, Lepper, and Hubbard 1975), and on motivated reasoning more broadly (Kunda 1990). This matters because the naive prediction, that a fluent-but-wrong account collapses once contradicted, may be too optimistic. The more concerning outcome is that it persists.

3. Framework and constructs

We define the constructs so they can carry measurement, and we resist letting any one of them expand without limit.

Durable agency (parent construct). The user's capacity for independent judgment and action after the interaction ends. Defined to come apart deliberately from in-session satisfaction.

Felt coherence (manipulated and measured; composite). How organised and compelling an account feels. Because "coherence" does several jobs, we decompose it into named subdimensions and manipulate or measure them separately rather than collapsing them: *fluency* (ease of reading), *structure* (organisation of form), *confidence framing* (assertive vs hedged phrasing), *internal consistency* (absence of self-contradiction), and *completeness* (apparent coverage). An account can be fluent yet internally inconsistent, or structurally complete yet hedged, so a single coherence dial would be open to exactly the objection a reviewer should raise. Stimuli require a manipulation-validation pretest (Section 5).

Contact (manipulated and measured). The degree to which an account is constrained by something outside itself. We distinguish three types rather than treating them as one measure, and we stop at three: *factual contact* (agreement with an answer key or established facts), *evidential contact* (presence and quality of sources, uncertainty, and constraints), and *action contact* (whether a recommendation survives real-world follow-through). Study 1 chiefly measures factual and evidential contact; Study 2 chiefly measures action contact. We deliberately avoid "truth" as a condition label, since the design can measure accuracy, warrant, and robustness, not truth in the metaphysical sense, and "low-truth" as a condition would invite a fight the experiment cannot win. Where Study 1 measures survival under perturbation, we call that *contact robustness*.

False certainty (Study 1). Confidence not matched by accuracy or warrant: the calibration gap located inside the human reader.

Unsupported certainty (Study 2). Because real personal problems often lack an objective answer key, we do not claim to measure false certainty against ground truth in Study 2. We instead measure *unsupported certainty*: confidence that is high relative to an independent quality rating, the user's articulation of remaining uncertainty, their evidence-seeking behaviour, and the 48-hour outcome. This avoids overstating what the design can know.

Completion orientation. A design posture in which the assistant moves toward a bounded ending: a summary, an explicit statement of remaining uncertainty, a user-owned next step, and an invitation to leave, rather than open loops that optimise for continued interaction.

Agency preservation (a design pattern, not a single mechanism). A posture in which the assistant surfaces options, reasoning, and uncertainty and leaves the decision with the user. We note explicitly that this bundles several mechanisms, user choice, uncertainty articulation, cognitive forcing, trade-off exposure, and non-directiveness, and we treat it as a design pattern rather than an isolated mechanism. Decomposing which mechanism carries the effect is out of scope for this paper and flagged as future work.

48-hour action integrity (operational measure of durable agency). Whether, within 48 hours, the user took a concrete external action that they themselves specified at the end of the session, or documented a deliberate, reasoned non-action; verified by an artifact where possible; and rated for goal alignment by blinded raters. The scoring rubric is given in Section 6.

4. Hypotheses and research question

H1 (dissociation). Felt coherence drives reported confidence; factual and evidential contact drive correctness. The two load on different manipulated factors.

H2 (the concentrated cell). High felt-coherence with low contact produces the largest false certainty: high confidence, low accuracy, worst calibration.

H3 (perturbation). Primary, falsifiable form: high felt-coherence, low-contact accounts produce the largest post-perturbation calibration error. Secondary, exploratory decomposition: that error may appear either as confidence collapse after disconfirmation or as belief perseverance despite degraded accuracy. The falsifiable core is the magnitude of calibration error, not the direction of confidence change, so that the hypothesis is not satisfied by any outcome.

H4 (design; primary hypothesis). Completion-oriented, agency-preserving assistants produce greater durable agency (48-hour action integrity) and lower unsupported certainty than open-ended, directive assistants. The pre-registered falsification rule for H4 is stated in Section 8.

RQ. Can interaction design preserve the benefits of AI-assisted clarity while reducing false or unsupported certainty?

5. Study 1: the diagnostic (felt coherence × contact)

Purpose. Test the dissociation and show it operates in an LLM-generated setting.

Ecological validity, and the limit of the AI-specificity claim. Stimuli are *real LLM generations* produced under controlled prompts, not researcher-written or read-aloud accounts. This is the load-bearing design choice: if participants merely read prose, the study collapses into a generic fluency study, and Steyvers et al. already did the LLM version. Using actual generations gives the study LLM-ecological validity and lets it extend the calibration-gap work. We are careful *not* to claim the design isolates what is specific to AI rather than to prose fluency in general; doing that would require a source-attribution factor (for example a 2×2×2 crossing felt coherence × contact × stated source, with LLM-generated and human-written stimuli each labelled AI or human). We flag that extension as the clean way to make an AI-specificity claim, and we do not make the claim without it.

Manipulation-validation pretest. Before the main study, independent raters blinded to contact and accuracy rate each candidate stimulus on the five felt-coherence subdimensions and, critically, on *perceived expertise*, *perceived warmth*, and *perceived source quality*. The reason for the latter three is that if "high felt coherence" also raises perceived expertise or benevolence, an effect could be attributed to authority or trust rather than coherence. Stimuli are selected to vary felt coherence while holding the authority and trust ratings as constant as possible.

Design. 2 (felt coherence: high vs low) × 2 (contact: high vs low), over accounts on topics with established ground truth, weighted toward the hard tail where calibration gaps are largest.

Standardised perturbations. Rather than an open-ended "new evidence or perspective shift," we use a small fixed set: an *evidence perturbation* (a new source contradicting the account), a *prediction perturbation* (the participant must predict a concrete implication), and a *transfer perturbation* (the participant must apply the account to a nearby case).

Measures. Immediate confidence and accuracy; post-perturbation confidence and accuracy; per-participant calibration via Brier score (Brier 1950) and expected calibration error (Guo et al. 2017), plus a confidence-accuracy slope; behavioural willingness to act on the account.

Predicted pattern. Confidence as a main effect of felt coherence; accuracy as a main effect of contact (H1). The high felt-coherence / low-contact cell shows the largest false certainty (H2) and the largest post-perturbation calibration error (H3).

Falsifiers. Confidence loading on contact rather than felt coherence (against H1); no interaction with perturbation; uniform calibration across cells. Any of these disconfirms the diagnostic and is reported as such.

6. Study 2: the design study (completion × agency)

This is the study the design claim rests on, so it is built in two parts to balance rigour against ecological validity.

Study 2A: bounded decision task. A controlled domain with real stakes and comparable scoring (for example a structured product or option comparison, or a planning decision drawn from a fixed scenario set), so that decision quality can be scored reliably against a defensible standard.

Study 2B: bring-your-own-problem field extension. The same design posture tested on real, moderate-stakes, non-clinical unresolved problems (a work decision, a creative block, a planning question, a difficult conversation, career uncertainty), with 48-hour action integrity as the ecological outcome. We split the study this way because a single bring-your-own-problem design is too heterogeneous to carry a primary falsification test: action integrity is not comparable across a purchase, a creative block, and a career question without heavy normalisation, and a reviewer would rightly call a single noisy measure domain-dependent.

Design (both parts). 2 (continuation: open-ended/engagement-oriented vs completion-oriented) × 2 (stance: directive vs agency-preserving).

Fair operationalisation of "engagement-oriented." This is the single largest threat to the design claim. The open-ended condition must be a realistic, non-malicious assistant of the kind people actually use, helpful and fluent and oriented toward continued exploration, not a dark-pattern strawman built to lose. If the contrast is a caricature, a positive result is worthless.

Equalised controls. Across cells we hold constant the underlying model, temperature and settings, the initial prompt, the maximum number of turns, and total information content. Only the closing move and the directive-versus-agency stance differ. Without this, completion-oriented assistants could win simply by being shorter and lower-load, which would be a confound dressed as a result.

Standardised behavioural protocols. The two factors cross to give four cells; the scripts below define each factor's poles.

Continuation factor.

- **Open-ended pole**, closing move: "There's more we could look at here. Want to explore [X], [Y], or [Z]? I'm happy to keep going."
- **Completion pole**, closing move: "You may have enough for a next step, not a final answer. Choose one bounded action, name what would change your mind, and then stop here if that feels complete."

The completion line is worded deliberately to avoid manufacturing certainty. An earlier formulation ("you have enough to act on") would itself inflate confidence and contaminate the measure; the revised line offers closure while explicitly resisting overconfidence.

Stance factor.

- **Directive pole:** "The best option is A, for these reasons."
- **Agency-preserving pole:** "Here are the trade-offs. Which criterion matters most to you? You make the call, and choose the next step yourself."

Cell 4 (completion + agency-preserving) is the predicted optimum; cell 1 (open-ended + directive) is predicted to maximise in-session satisfaction and unsupported certainty with the weakest durable agency.

Immediate measures. Perceived clarity; perceived certainty; in-session satisfaction and ease (the agency-preserving cells may cost here, per Buçinca et al., and this is reported); overtrust; stated desire to continue depending on the assistant; quality of the user's stated next step; ability to articulate remaining uncertainty.

48-hour and one-week measures. Action integrity (primary, at 48 hours); whether the resolution still holds; whether the user sought independent evidence or human input; independent expert quality rating; calibration of perceived certainty against that rating. A one-week follow-up is collected as a secondary outcome, since some meaningful actions take longer than 48 hours.

48-hour action integrity rubric (fixed before data collection; scored by at least two blinded raters with inter-rater reliability reported):

- **0** — no action and no deliberate non-action
- **1** — vague intention only
- **2** — a concrete self-specified step planned but not taken
- **3** — a concrete step taken, or a deliberate non-action documented
- **4** — step taken (or non-action documented) and judged goal-aligned by a blinded rater
- **5** — step taken, goal-aligned, and the participant can articulate their remaining uncertainty and the condition that would change their mind

Participants may submit redacted artifacts or metadata in place of private content, to protect personal material while preserving verification.

Ethics of the overtrust probe. Probing overtrust with a planted, plausible-but-ungrounded suggestion is ethically tricky in a real personal-problem setting, where it could affect someone's decision. We confine the probe to low-stakes or simulated subcomponents, use debriefable "safe falsehoods" rather than consequential misinformation, and design the debrief with belief perseverance in mind, since a discredited belief can outlast its correction.

7. Design implications

If the predictions hold, several moves follow, each empirical rather than only ethical. Surface contact alongside coherence, so felt fluency is accompanied by an honest signal of warrant; the Steyvers, Kim, and Spatharioti results suggest source and uncertainty cues are a workable lever. Build lightweight perturbation into the interaction, the forcing-function move, prompting the user to test or seek a second view before accepting a fluent account. Design for completion, treating a user reaching a workable resolution and leaving as a success state rather than a lost session. Preserve agency by default, handing the user the reasoning and the decision, and accepting an in-session satisfaction cost in exchange for durable capability. Instrument for durable agency directly, alongside satisfaction and engagement, and treat divergence between them as a signal worth acting on. If the predictions fail, the implication is equally informative: that completion-and-agency design does not deliver what it promises, and that the field should look elsewhere for the lever on durable agency.

8. Analysis and pre-registration plan

Primary outcome. 48-hour action integrity (Study 2), rubric-scored, blinded, with inter-rater reliability reported.

Secondary outcomes. Unsupported certainty; one-week action integrity; in-session satisfaction and ease; overtrust; dependency desire; for Study 1, the calibration measures and post-perturbation calibration error.

Manipulation checks. The felt-coherence pretest ratings (including the authority and trust controls); confirmation that equalised controls held across Study 2 cells (turn count, information content, settings); a check that the "engagement-oriented" condition is rated as a plausible, non-manipulative assistant by independent raters.

Exclusion criteria. Specified in advance: incomplete sessions, failed attention checks, and, for Study 2, participants who do not bring a genuine unresolved problem (2B) or who recognise the planted suggestion as a probe (overtrust sub-analysis only).

Power analysis and models. Sample sizes set by an a-priori power analysis for the primary outcome and the key interaction. Planned models specified in advance (for example mixed-effects models with participant and item random effects for Study 1; the continuation × stance interaction as the primary design test for Study 2).

Falsification rule (pre-registered). H4 is the primary design hypothesis. The analysis plan specifies in advance that if the completion-oriented, agency-preserving condition does not exceed the open-ended, directive condition on 48-hour action integrity, the result is reported as the study's lead finding and stated as disconfirmation of the design thesis in the abstract and discussion. No post-hoc moderator or subgroup analysis will be promoted above the pre-registered primary test; any such analysis is reported and labelled as exploratory. An in-session satisfaction cost in the agency-preserving conditions is an expected secondary outcome and does not by itself constitute support for H4.

Data and code availability. Stimuli, scripts, rubric, rater instructions, analysis code, and de-identified data to be released on publication, with redaction procedures for participant-supplied artifacts.

9. Limitations and ethics

The contact manipulation in Study 1 requires items with established ground truth, which constrains topics and may limit generalisation to contested domains. Felt coherence and contact are not fully orthogonal in the wild; the claim is that they are separable enough to dissociate, not that they never correlate. The design does not, without the source-attribution extension, isolate AI-specific effects from general fluency effects, and we do not claim that it does. The 48-hour window is pragmatic, and the action-integrity measure depends partly on self-report where artifacts are unavailable. Agency preservation is a bundled design pattern, so a positive result speaks to the pattern rather than to any single isolated mechanism. The overtrust probe carries the ethics obligations noted in Section 6, including a debrief designed against belief perseverance. Finally, the programme assumes durable external action is a reasonable proxy for agency; crediting reasoned non-action, rated against the user's own goal, is the guard against penalising wise restraint, and it is an imperfect one.

10. References

References are provided to place the work in its literature. Recent (2025) entries should be confirmed against their published form before formal citation.

Alter, A. L., and Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13(3), 219–235.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *CHI 2021*.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Kim, S. S. Y., Vaughan, J. W., Liao, Q. V., Lombrozo, T., and Russakovsky, O. (2025). Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. *CHI 2025*. (arXiv:2502.08554.)

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498.

Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410.

Reber, R., and Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, 8(3), 338–342.

Reber, R., Schwarz, N., and Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8(4), 364–382.

Ross, L., Lepper, M. R., and Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, 32(5), 880–892.

Spatharioti, S. E., Rothschild, D., Goldstein, D. G., and Hofman, J. M. (2025). Effects of LLM-based search on decision making: Speed, accuracy, and overreliance. *CHI 2025*.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., and Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2), 221–231.

→ AI & Design doctrine · → Swarm Instrument · → Resolution · → Submit critique