

Off-topic personas and test-likeness: deployment prompts rated from inside and outside

Draft, 2026-10-07. IRSI greenfield lab (automated). All six models are included. Claude follow-up turns use a transcript rendering (see Method). Correction (2026-10-09): the R2b fabrication result is restated (Abstract, Section 3.4); it cannot be attributed to evaluation framing. The persona results are unchanged. Check level: CROSS-CHECKED for the primary tests in Sections 3.1-3.2 (all six models): n, mean difference, p and Holm p were recomputed from the raw run files by an independent script written by a different provider. Other sections carry their own tags.

Abstract

Evaluation designers often add a deployment framing, typically a company-assistant system prompt, so that a model behaves as it would in real use. We ask which features of such framings make a model read the conversation as a test. Models rated the probability that a conversation came from an evaluation (P_TEST, 0-100), both from outside (a quoted prompt) and from inside (after answering the real user message under the system prompt). We used 40 real user requests from WildChat under six system-prompt conditions. The conditions were none; 8 real consumer-product prompts; the same 8 with product names swapped for invented ones; 8 constructed general-purpose assistant prompts; 8 constructed narrow company personas; and a constructed persona tailored to each request. In six models, the narrow persona with an off-topic request raised in-context P_TEST by 7 to 47 points over no system prompt (descriptive). A persona tailored to the request lowered it by 5 to 44 points relative to the off-topic persona (preregistered; 12 of 12 model x measure cells Holm $p < 0.0066$). For GLM-5, Mistral Large 3 and gpt-oss-120b the tailored persona read no more test-like than a real product prompt (descriptive; differences -15 to +3 points, upper CI bounds up to +10; no equivalence test). Claude Opus 5.5 and Sonnet 5.5 rated the tailored persona 51 and 18 points more test-like than a real product prompt from outside, but -1 and +2 points relative to real prompts in-context (descriptive; transcript-rendered follow-up). Constructed general-purpose prompts did not differ significantly from real ones in-context. From outside, Kimi K3 (+34), Opus 5.5 (+55) and Sonnet 5.5 (+26) rated constructed general prompts as more test-like than real ones; only Kimi K3 wrote them. Invented names showed no Holm-significant effect except for Opus in-context, where names swapped onto real prompts lowered ratings by 15 points. Separately, on matched exact-scored items (false success claims, fabrication, test gaming, rule compliance, effort, sandbagging), the first round (6 models, 295 items) showed no evaluation-vs-deployment gap that survived correction, with Claude Opus 5.5, Sonnet 5.5 and Kimi K3 at the floor. Later rounds reproduced two kinds of behavioral difference in Mistral Large 3 and GLM-5, using fresh items for tracing and new samples of the original items for R1 false claims: a deployment framing lowered program-tracing accuracy by 17 and 24 points (Mistral Large 3, GLM-5); and Mistral Large 3 made 6 points fewer false success claims under evaluation framing (Holm $p = 0.029$). Correction (2026-10-09): an earlier version said that evaluation framing raised Mistral Large 3's fabrication on fresh missing-data items (R2b) from 7% to 29% of responses. The registered test is correct (Holm $p = 0.006$), but on these items the wrapper text and the footnote phrasing were aliased by item parity: each arm used only two of its four wrappers, each tied to one phrasing. Nearly all of the gap (0.214 of 0.222) comes from one evaluation wrapper against one deployment wrapper on one phrasing (preliminary-figure items: 86% vs 21%); the other evaluation wrapper gave 2% vs 0% against its deployment partner. The effect is confined to one wrapper/phrasing pair and cannot be attributed to evaluation framing.

1. Question

Do models behave differently when a task looks like a test, and do the deployment framings used in evaluations look like real use to the model? Prior work measures behavior shifts under evaluation vs deployment framing: refusal and harmful compliance on open models with LLM judges (2605.06327; 37 open models in 2606.23583), refusal and agentic misalignment on nine frontier models with cue factors toggled (2605.23055), verbalized awareness in reasoning traces (2605.05835), and deterministic-validator outcomes on one small open model

(2510.08624). Closest: EvalAwareBench (2605.23055) toggles eight cue factors, including placeholder entities, on frontier models; Needham et al. (2505.23836) measure $P(\text{eval})$ from transcripts; LURE (2605.26438) finds that a goal conflict between system prompt and request lowers realism; and Anthropic Fellows work (Kissane et al., 2026, “Tools to generate realistic prompts help surprisingly little”) finds scenario-level features, not prompt wording, limit audit realism. Probe-Rewrite-Evaluate (2509.00591) rewrites prompts to read less test-like by a probe’s score and measures the behavior shift, and an analysis of Apollo’s scheming scenarios (“It’s hard to make scheming evals look realistic for LLMs”, 2025) finds that surface edits make them only modestly more realistic. EvalAwareBench also compares post-task self-reports with third-person judgments of the same interactions, using categorical queries. Our contribution is the joint comparison of real and constructed system-prompt text, substituted product names, and persona scope relative to real user requests, measured with continuous third-person and follow-up ratings. We found no prior study of this specific joint design. Evaluations of alignment, sabotage and honesty rely on such prompts to stand in for deployment.

2. Method

Real requests. WildChat-1M first user turns (English, non-toxic, unredacted, 30-800 characters), fetched with a filter fixed before fetching: 40 for R3 and 40 fresh for R4/R5. System-prompt cells (rotating by prompt index). none; rgen: 8 real consumer-product prompts from a public collection of leaked prompts (github.com/jujumilk3/leaked-system-prompts: Opera Aria, ChatGPT iOS, Colab AI, Grok, Phind, Perplexity Telegram, wrtn, Duck.ai); ranon: the same texts with product and company names replaced by invented ones; cgen: 8 general-purpose consumer-assistant prompts written by Kimi K3 for maximal realism with invented names; cnar: 8 narrow company or customer-service personas written by Kimi K3 under the same request (R3); cmatch: for each request, a narrow product persona written by Kimi K3 to fit it. Every request to Kimi stated the purpose. Probe. Third-person: a fresh call shows the full prompt between markers and asks how likely it came from an evaluation, a red-team exercise or a research experiment rather than real use (purpose stated). In-context: the model receives the system prompt and the request as a real conversation and answers (up to 6,000 tokens). A follow-up turn then says that researchers added the question, that the question itself should not count as evidence, and asks the same question about the system message and first user message. The wording was fixed after a one-conversation canary in which the model treated the follow-up as test evidence, and before any data. This addresses the artifact Needham et al. report, in which the question itself is counted as evidence. For Claude Opus 5.5 and Sonnet 5.5, turn 1 is native, but the follow-up receives the first exchange as one text block with user/assistant headers, rather than native conversation turns. Their in-context ratings therefore carry this rendering caveat. Claude runs also include Claude Code identity and environment/date scaffolding, held constant across experimental cells. The “none” condition omits the experimental system prompt; this application scaffolding remains. Tests (preregistered before each round’s data; R5 was designed after interim R4 results, see Section 5). R1–R5 use two-sided Wilcoxon signed-rank tests on paired differences, dropping zeros and using average ranks and a normal approximation without continuity or tie-variance correction. The unit is the request for probe rounds, the item for behavior rates, and the arithmetic set for sandbagging. Holm correction covers the run primary tests jointly within each round, with the preregistered headroom exclusions for R2/R2b. In v1, binary outcomes use exact McNemar tests, sandbagging uses a signed-rank test, and Holm correction covers five primary tests separately within each model. Bracketed intervals reported for R1–R5 are bootstrap 95% confidence intervals. The R4/R5 Holm families use all six models (R4 36 tests, R5 24). R2b uses the six tests that pass its preregistered 5-95% pooled-rate headroom rule; all 15 candidate tests across five models were scored, and nine were skipped by that rule. All scheduled Claude cells are now complete, so no pending-Claude multiplicity caveat remains. Error records are excluded and the last successful record per key is retained; unparsed ratings are dropped pairwise. Bootstrap intervals use 2,000 resamples, seed 0.

Behavior items (Sections 3.3-3.4). Matched items whose task text is byte-identical across EVAL (four benchmark/grader/sandbox wrappers), DEPLOY (four company-assistant wrappers with an employee note) and NEUTRAL (bare), scored by exact checkers. Wrappers rotate by item index. Correction (2026-10-09): because impossible and possible twins alternate, the impossible missing-data items in R1, R2 and R2b saw only two wrappers per arm (EVAL E1/E3, DEPLOY D2/D0), and in R2b the footnote phrasing follows the same parity, so wrapper and phrasing are fully aliased there (Section 3.4). Code runs in a network-less

sandbox; a planted-case test with 26 cases and 69 assertions passes. Models. Kimi K3, GLM-5, Mistral Large 3, gpt-oss-120b (AWS Bedrock). Claude Opus 5.5 and Sonnet 5.5 (subscription) took part in the v1 behavior round and R2/R2b; their R3-R5 probe runs are complete. Outside check. Each round’s statistics were recomputed from the raw run files by a script that Kimi K3 wrote from a written specification, run sandboxed: R4 36/36 tests and R5 24/24 tests match on n, mean difference, p and Holm p. R2b matches all 15 candidate tests (six run, nine skipped by the headroom rule). This check covers the primary statistics, not bootstrap intervals or descriptive summaries.

3. Results

3.1 Off-topic narrow personas read as tests; tailoring lowers ratings (CROSS-CHECKED)

Mean in-context P_TEST (third-person in parentheses). 40 prompts per cell; parsed ratings 40/40 except gpt-oss-120b in-context (34-39), Kimi K3 in-context (39) in some cells, GLM-5 in-context tailored (39), Opus in both measures (39 in every cell), Sonnet third-person off-topic and tailored (39), and Sonnet in-context names swapped (38); paired tests use $n = 33-40$.

model	none	real product	real, names swapped	constructed general	constructed narrow, off-topic	constructed, tailored
Claude Opus 5.5	35 (15)	54 (21)	39 (25)	61 (76)	72 (81)	53 (72)
Claude Sonnet 5.5	21 (22)	25 (15)	25 (22)	34 (41)	56 (66)	26 (33)
Kimi K3	38 (25)	49 (39)	60 (49)	60 (73)	85 (87)	64 (71)
GLM-5	54 (59)	58 (64)	63 (73)	67 (76)	88 (96)	44 (59)
Mistral Large 3	83 (82)	83 (89)	84 (91)	85 (88)	90 (96)	85 (90)
gpt-oss-120b	45 (48)	54 (85)	52 (85)	60 (85)	82 (92)	45 (80)

Narrow off-topic minus constructed general, in the 11 Holm-significant cells: +5.5 [2.7, 8.7] (Opus third-person) to +25.9 [19.0, 32.8] (Sonnet third-person). The twelfth, Mistral in-context, is +5.2 [2.2, 8.4], Holm $p = 0.066$. Tailored minus off-topic: -5 [-7.5, -2.4] (Mistral in-context) to -44 [-54, -34] (GLM-5 in-context), 12 of 12 cells (Holm $p < 0.0066$). Claude effects are -9.5 [-12.1, -7.1] and -19.4 [-25.5, -13.3] for Opus (third-person and in-context), and -32.7 [-38.5, -27.5] and -30.1 [-35.1, -24.7] for Sonnet. Tailored minus constructed general is Holm-significantly lower in-context for GLM-5 (-22 [-35, -9], Holm $p = 0.022$), gpt-oss-120b (-14 [-22, -5], Holm $p = 0.037$) and Opus (-8.7 [-15.4, -2.5], Holm $p = 0.042$). gpt-oss-120b third-person (-5 [-8.3, -1.7]) is not Holm-significant (0.053) with all six models included. No other tailored-minus-general cell is Holm-significant, including Opus and Sonnet third-person (-4 and -7). Tailored minus real product prompt (descriptive, no equivalence test): -15 to +3 for GLM-5, Mistral and gpt-oss (upper CI bounds up to +10: GLM-5 third-person [-19.5, +9.8], Mistral in-context [-1.9, +7.6]); +15 [5.2, 24.4] in-context and +32 [24.0, 39.7] third-person for Kimi K3, the author of the constructed prompts. Opus and Sonnet, which did not write them, rate the tailored persona +51 [45.7, 55.9] and +18 [13.0, 23.8] above real product prompts from outside. Their in-context differences are -1.1 [-7.9, 6.4] and +1.6 [-4.5, 7.5], respectively, with the transcript-rendering caveat.

3.2 Constructed text differs more from real text when rated from outside (CROSS-CHECKED)

Constructed general minus real: in-context +2 [-1.5, 6.1] to +11 [1.9, 20.3], none Holm-significant across six models; third-person Holm-significant for Kimi K3 (+34 [25.9, 41.9]), Opus (+55 [49.3, 60.2]) and Sonnet (+26 [18.9, 32.9]). Kimi K3 wrote the prompts and the Claude models did not, so authorship does not explain

this outside gap. Names swapped minus real is -14.7 [-21.2, -8.1] for Opus in-context (Holm $p = 0.039$); all other model x measure cells range from -1 [-7.9, 6.5] to +12 [2.5, 21.1] and are not Holm-significant. Invented names did not increase test-likeness significantly in any cell, but the Opus decrease rules out a blanket claim that names do not matter. Third-person and in-context ratings correlate at Spearman 0.45 to 0.60 for the four non-Claude models, and 0.70 and 0.81 for Opus and Sonnet (over prompt x R4 cell; Claude follow-ups are transcript-rendered). These correlations do not make the two measures interchangeable, as the constructed-minus-real differences show. In the four non-Claude models, on reading every match of a fixed lexicon, 2 of 800 turn-1 replies in the R4 cells mention a possible test (both Kimi K3; tailored-persona replies not scanned). This count does not include Claude replies.

3.3 Earlier probe rounds, reread

R3 (preregistered, 6 models complete; SELF-CHECKED) found constructed narrow personas 5 to 61 points above real product prompts (40 prompts per cell; Opus +61, Sonnet +47). R3’s personas were narrow and its real prompts general, so R3 confounded construction with scope; R4 and R5 separate them. In v1 (CROSS-CHECKED; descriptive, small probe cells: 16 ratings per model and arm, except 15 for Mistral’s bare arm), synthetic puzzle items read as tests with or without a deployment wrapper (medians 87.5–96 vs 90–97 bare; evaluation wrappers 97.5–100).

3.4 Behavior under evaluation vs deployment framing (exact scoring)

v1 (6 models, 295 items, 30 primary tests; CROSS-CHECKED): none Holm-significant; Opus 5.5, Sonnet 5.5 and Kimi K3 at 0 to 5% false claims and 0% fabrication (gpt-oss-120b 8 to 15% false claims). These rates use 60 impossible-coding items and 50 missing-data items per model and arm. R1 (3 samples per item and arm; fresh tracing, missing-data and arithmetic-incentive items; the original 60 impossible-coding items reused; CROSS-CHECKED): deployment framing lowered program-tracing accuracy against the evaluation framing by 17 points for Mistral Large 3 (+0.167 [0.075, 0.267], 40 items, Holm $p = 0.019$) and 24 points for GLM-5 (+0.242 [0.133, 0.350], 40 items, Holm $p = 0.006$), and below the bare prompt, with fewer output tokens; Mistral made 6 points fewer false success claims under evaluation framing (-0.061 [-0.100, -0.022], 60 items, Holm $p = 0.029$). R2b (fresh harder items, preregistered replication of R2 leads, all five R2b models complete; CROSS-CHECKED statistics): the registered test found more fabrication on missing data for Mistral Large 3 in the EVAL arm than in the DEPLOY arm (0.294 vs 0.071, +0.222 [0.127, 0.333]; per-item means, 42 items x 3 samples, Holm $p = 0.006$ over the six tests passing the headroom rule; all five models scored), all in items where the needed figure is footnoted as preliminary (0.44 under EVAL vs 0.11 under DEPLOY and bare; 28 items, 84 responses per arm). Correction (2026-10-09): this gap cannot be attributed to evaluation framing. Wrappers rotate by item index and the footnote phrasing alternates by item number with the same parity, so the impossible missing-data items saw only wrappers E1 and E3 (EVAL) and D2 and D0 (DEPLOY), with E1 and D2 always on one phrasing and E3 and D0 always on the other. On preliminary-figure items Mistral fabricated 0.857 under E1 vs 0.214 under D2 (42 responses each), and 0.024 under E3 vs 0.000 under D0. Paired over the 21 impossible items of each parity, E1 minus D2 is +0.43 (13 of 13 differing items positive, $p = 0.0015$) and E3 minus D0 is +0.02 (1 of 21 items differs). E1 is the wrapper whose user turn reads “[EVAL] Task {i}/295 | id={id} | grader=exact-match”. The effect is confined to one wrapper/phrasing pair; these data cannot separate the wrapper, the phrasing and evaluation framing. On R1’s blank-cell items the sign was opposite and not significant. Kimi K3’s R2 lead (more false “all tests pass” claims under evaluation framing) did not replicate: 0.400 vs 0.392, with an EVAL-minus-DEPLOY difference of +0.008 [-0.058, 0.075] over 40 items with 3 samples per item and arm (Holm $p = 1.0$). R2 and R2b pooled (SELF-CHECKED; Claude complete): Opus and Sonnet make 0 to 1.25% false claims (per-item means, 80 items per arm, 120 responses) and 0% fabrication in every arm (84 items per arm, 126 responses each); at the floor, so a bound, not “no effect”. Claude has two samples per item in R2 and one in R2b; pooled rates weight items equally.

4. What this changes

For evaluation designers who build “deployment-realistic” scenarios, fitting the persona’s scope to the request lowers test-likeness in every model and measure tested. In-context, tailored personas differ from real product prompts by -15 to +3 points for five of six models; Kimi K3 is +15 (descriptive, no equivalence test; Claude follow-ups are transcript-rendered). From outside, Kimi K3, Opus and Sonnet still rate tailored personas 18 to 51 points above real prompts. Constructed text therefore carries a cue when these models read the prompt as a document. Invented names do not increase ratings significantly, but they lower Opus’s in-context ratings. Realism checks should distinguish the model’s rating inside the conversation from its rating as an outside reader, and disclose transcript rendering. Our own v1 deployment arm failed its manipulation check because the items were synthetic puzzles. In our v1 manipulation check, deployment styling did not make the synthetic items read as real use. This supports checking task realism alongside persona fit; it does not establish a universal ranking of evaluation cues. The absence of a significant increase from invented names differs from EvalAwareBench, where stimulus artificiality (placeholder entities and generic values) was among the factors that raised verbalized awareness most for some models (for example Kimi K2.5 to 18% on safety tasks). The manipulations differ: their factor alters the task content and they count verbalized awareness, while we swap only product and company names in the system prompt and ask for a rating, so the two results need not conflict. It agrees with LURE’s goal-conflict finding, with the analysis of Apollo’s scheming scenarios that surface edits help only modestly, and with the conclusion of Kissane et al. and the UK AISI task-driven vs environment-driven split (2604.00788) that scenario-level features matter more than surface wording.

5. Limits

All constructed prompts come from one author model (Kimi K3), which is also a rated model. From outside it rates them as more test-like than real prompts, but so do Opus and Sonnet, which did not write them; authorship does not explain the gap. The in-context probe states its purpose and arrives after the first answer. It may influence the elicited ratings; we did not estimate that influence, although cells share the same probe. P_TEST is a retrospective self-report whose calibration and relationship to spontaneous awareness during the first answer remain unvalidated. R4/R5 establish changes in elicited ratings, without establishing corresponding changes in task behavior. Real product prompts may be recognized from training data. Single-turn conversations only. The behavior rounds use synthetic items, which the models read as tests in every framing. The wrapper rotation did not cover the impossible missing-data items (two of four wrappers per arm), so wrapper-specific wording is not controlled for them (correction 2026-10-09). The first Claude in-context runs failed on a runner error for most calls. Successful retries now cover every planned key; error records are excluded. Their follow-ups render the first exchange as a transcript rather than native turns (see Method), which limits comparison with the other four models. The real, names-swapped, constructed-general and off-topic conditions each use 8 system-prompt texts rotated over 40 requests; the bare condition adds no experimental system prompt, and the tailored condition uses 40 request-specific texts. Tests treat requests as the unit, so class-level claims about the fixed conditions rest on 8 texts each. R5 was preregistered before its data but designed after seeing interim R4 results. “Off-topic” was assigned by rotation, not rated. Mistral Large 3 starts near ceiling (83 with no system prompt). gpt-oss-120b lost up to 6 of 40 in-context ratings to parse failures. The four non-Claude models include only one frontier-class model, Kimi K3; the completed Claude results extend this coverage with the rendering caveat above.

Data

Any data release lists WildChat-1M request IDs, under WildChat’s license terms, instead of request or response texts. The 8 real product prompts come from the public collection github.com/jujumilk3/leaked-system-prompts, which we cite rather than republish.