DO LLMS KNOW HOW A PERSON HANDLES EVERYDAY SELF-CONTROL CONFLICTS? A PERSON–EPISODE TEST WITH EXPERIENCE SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-control in daily life is a process: when people face a conflict, they select strategies that depend on who they are and on the situation, appraise how well they managed, and monitor and evaluate their strategy use. Large language models (LLMs) match human performance on several social-inference tests and are increasingly used to model and support people, yet their evaluations score outcomes, such as keyed answers, group-level rates, and questionnaire scores, leaving open whether a model reproduces the process by which a specific person handles a specific episode. We test 22 LLMs, prompted in English and German with increasingly detailed person descriptions, on 3,725 experience-sampled self-control conflicts from 468 participants, and score each predicted step (the person's strategies, success, and metacognition) against the person's own reports and reference predictors. At the selection step, the models list more strategies than people report, predict similar sets for different people, and recover the reported strategies less well than a rule that lists the same number of the most frequent strategies. At the appraisal step, person information improves their ranking of people by typical success and changes their tracking of a person's episodes only slightly. At the metacognitive step, their ratings follow reported success and strategy counts, ranking metacognitive knowledge weakly and missing regulation, which the conflict histories carry. The models thus capture what people generally do and how people differ, and miss how a person's situation shapes strategy choice, how strategies shape the appraisal of an episode, and how strategy use across episodes reflects metacognition. Because each step looks plausible on its own, evaluations of LLMs as models of individuals should test this process directly.

1 INTRODUCTION

People face self-control conflicts every day, when they must start an aversive but important activity, continue it, or resist a temptation, and handle them through a process: they select strategies, often several at once, appraise how well they managed, and plan, monitor, and evaluate their strategy use (Hofmann et al., 2012; Milyavskaya et al., 2021; Hennecke & Bürgler, 2023). Each step depends on the person and the situation (Wenzel et al., 2024; Bürgler et al., 2021), and the steps relate to the outcome: using more strategies and reporting more metacognitive knowledge, planning, and monitoring are associated with higher success (Bürgler & Hennecke, 2024; Bürgler et al., 2022). Self-control in daily life is thus a process specific to the person and the episode.

Large language models (LLMs) match human performance on several tests of social inference, such as inferring others' false beliefs and choosing the most effective way to manage an emotion (Kosinski, 2024; Strachan et al., 2024; Schlegel et al., 2025), and are increasingly used as models of human cognition and of individual people (Binz et al., 2025; Liu et al., 2026) and as coaches for behavior change (Jörke et al., 2025). These uses depend on how well an LLM reproduces how a person arrives at an action and an appraisal in a situation, the process that cognitive models explain and coaching aims to change, yet evaluations of LLMs score outcomes, such as keyed answers, agreement with group-level data, or questionnaire scores, mostly across people. A model can then supply plausible strategies and typical outcomes from general knowledge while misrepresenting the process, an error that matters most where the process varies within a person, as in daily self-control, where 82% of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the variance in success lies within persons (Bürgler & Hennecke, 2024), and that can carry into conclusions from LLM-simulated participants or LLM-based assessments (Wang et al., 2025). Tests of whether LLMs reproduce each step of a specific person's process remain limited.

We therefore test, step by step, whether LLMs reproduce this process for specific people, using an experience-sampling study in which 468 participants reported 3,725 self-control conflicts, the strategies they used, and how successful they felt, linked to their demographics, personality, trait self-control, and metacognition (Bürgler & Hennecke, 2024). We prompt 22 LLMs in English and German with increasingly detailed person descriptions to predict each step, and we score every prediction against the person's own report, which we call *person–episode fidelity*, and against reference predictors, the person's earlier reports, and human evaluators. We ask three research questions:

- **RQ1 (action):** Do LLMs recover the strategies a person reported in an episode?
- **RQ2 (person information):** Does richer information about the person improve the predicted strategies and success, across people and within a person?
- **RQ3 (metacognition):** Does a person's conflict history allow LLMs to rank people by metacognitive knowledge and regulation?

The results are consistent across the three steps: the models list about three times as many strategies as people report, predict similar sets for different people, and recover the reported strategies less well than a rule that lists the same number of the most frequent strategies (RQ1; Sections 4.1–4.2); person information improves their ranking of people by typical success and changes their tracking of a person's episodes only slightly (RQ2; Section 4.3); and their metacognition ratings follow reported success and strategy counts and miss regulation, which the histories carry (RQ3; Section 4.4). Selection thus follows a repertoire, appraisal follows the person, and metacognition follows outcomes: LLMs capture what people generally do and how people differ, and miss how a person's situation shapes strategy choice, how strategies shape the appraisal of an episode, and how strategy use across episodes reflects metacognition. As each step looks plausible, evaluations of LLMs in coaching, simulated participants, and psychological assessment, which have judged plausibility, group agreement, or questionnaire correlations, miss these failures; the results indicate that they should also test whether recommendations change with the person and the situation, whether appraisals follow the strategies used, and whether ratings hold beyond outcome summaries.

2 Related Studies

2.1 Self-Control in Daily Life

Experience-sampling studies show that people face self-control conflicts daily and resolve them with several strategies at once, 3.3 per conflict on average in the study that contains our data (Hofmann et al., 2012; Milyavskaya et al., 2021; Bürgler & Hennecke, 2024). Strategy use and success depend on the person and the situation together: the interaction of person and conflict type explained 34.3% of the variance in success and the specific strategy used 0.03% (Wenzel et al., 2024), and 82% of the variance in success lies within persons (Bürgler & Hennecke, 2024), so a prediction that is the same for all of a person's episodes can capture at most the remaining 18%. Metacognitive accounts add that people differ in knowledge about their self-control and in regulation, the planning, monitoring, and evaluation of strategy use (Hennecke & Bürgler, 2023); on the Metacognition in Self-Control Scale (MISCS), only Knowledge predicts success beyond trait self-control (Bürgler et al., 2022). Reproducing these steps for a particular person thus combines general knowledge of strategies with information about the person and the episode, a social inference on which LLMs match people in standard tests (Section 2.2); tests of this inference for actual individuals remain limited.

2.2 Social Cognition in LLMs

LLMs match or exceed human scores on social-cognitive tests with keyed answers, including false-belief tasks (Kosinski, 2024; Strachan et al., 2024) and emotional-intelligence tests such as choosing the most effective way to manage an emotion (mean accuracy 81% across five tests vs. 56% in their validation samples; Schlegel et al., 2025), and exposure to language alone yields some sensitivity to false beliefs (Trott et al., 2023). For individuals, LLM ratings of social-media posts correlate

Table 1: The four prediction tasks. An episode is given by its activity, goal domain, and conflict strength; success and strategies in the input are the participant's reported values.

Task	Step	Input	Output	Metric
1	Action	Persona, episode	Strategy set	Jaccard
2	Action given success	Persona, episode, success	Strategy set	Jaccard
3	Appraisal	Persona, episode, strategies	Success (1–7)	Absolute error, correlation
4	Metacognition	Demographics, conflict history	12 MISCS items	Spearman ρ

$r = .29$ with the writers' Big Five (Peters & Matz, 2024), persona prompts improve predictions little where persona variables explain little variance (Hu & Collier, 2024), and LLMs expect people to choose more rationally than they do (Liu et al., 2025), whereas a model fine-tuned on 10.7 million laboratory choices predicts held-out participants (Binz et al., 2025). These studies score outcomes, such as keyed answers, traits, questionnaires, or laboratory choices, mostly between people; tests of the process by which a specific person acts and appraises in everyday episodes, step by step and within and between persons, remain limited, and this study provides such a test.

3 Study Design

3.1 Experience-Sampling Data

Data sets. Testing person–episode fidelity requires three kinds of data from the same persons: reports of real everyday conflicts, the strategies used and the success felt in each, and person measures that can serve as persona information and as a metacognition target. All four tasks use a German experience-sampling study of daily self-control that provides all three (Bürgler & Hennecke, 2024), whose data Wenzel et al. (2024) also analyzed: 503 participants completed baseline questionnaires and then received eight prompts a day for ten days, and, of its two randomly presented questionnaire forms, we use the one that recorded strategies. Its baseline measures cover demographics and life status, Big Five personality (Soto & John, 2017), trait self-control capacity (Hoyle & Davisson, 2016), and the 12-item MISCS (Bürgler et al., 2022), validated in part on this sample, and its published analyses decompose the variance of success (Section 2.1), which bounds what a person description alone can predict. A U.S. experience-sampling study of food desires (686 episodes, 102 participants; Lopez et al., 2021) tests whether the Task 1 pattern generalizes (Section 4.2): it records multi-label strategy reports, the format Task 1 scores, in another country, domain, and strategy list.

Conflict reports. Each report concerns one of three conflict types, initiation (starting an aversive but goal-relevant activity), persistence (continuing it), or inhibition (resisting a temptation), and records the elements the tasks use: the conflict's strength (1–7), the strategies used (25 binary labels), and how well they managed and how satisfied they were (two 1–7 items, averaged into subjective success). The evaluation set contains 3,725 strategy-labeled episodes from 468 participants (1,909 initiation, 1,135 persistence, 681 inhibition; Appendix B.1). Tasks 1–3 are analyzed on persistence episodes, the type that records the activity being continued and its goal, two fields that give the model the demands of the episode; the other types are secondary checks (Appendix H).

Strategy groups. To describe the results, we group the 25 labels by content (Appendix Table 31): internal (e.g., self-talk, commitment), planning (e.g., goal setting), in-the-moment adjustments (changing feelings or the activity, distraction, suppression, substance use), environmental (e.g., removing temptations), social, goal-focused, training self-discipline outside the activity, and other.

3.2 Prediction Tasks

The tasks follow the process step by step (Table 1): selecting strategies in an episode (action), appraising how well one managed (appraisal), and knowing and regulating one's strategy use across episodes (metacognition). Each task gives the model part of the participant's record and scores its prediction of one step against the participant's own report, which locates each failure at its step.

Tasks 1–3. Task 1 tests the selection step (RQ1): from the persona and the episode, the model predicts the reported strategies, the inference behind uses of LLMs as models of individuals (Binz et al., 2025) and as coaches that tailor plans to a person (Jörke et al., 2025). Tasks 2 and 3 test the relation between selection and appraisal in both directions, an inference from the observed to the unobserved part of a person's experience that theory-of-mind tests probe with vignettes (Strachan et al., 2024): Task 2 adds the reported success to Task 1, so comparing the two shows whether knowing the outcome changes the inferred strategies and moves them toward the reported set, and Task 3 predicts the appraisal from the reported strategies. Because 82% of the variance in success lies within persons (Section 2.1), Task 3 is scored between and within persons, separating a stable success level from each episode's appraisal. The inputs are reported values (except 18 unrated Task 2 episodes; Appendix B.3), so each task isolates one step and errors do not carry over between steps.

Task 4. Task 4 tests the metacognitive step (RQ3): metacognitive accounts relate knowledge and regulation, rated by the participant at baseline, to how a person selects, monitors, and evaluates strategies across conflicts (Hennecke & Bürgler, 2023), so the model receives the demographics and the full conflict history (type, strength, strategies, and success of every conflict) and answers the 12 MISCS items, scored as for the participant. Inferring a disposition from behavior parallels LLM inferences of personality from social-media posts (Peters & Matz, 2024), and withholding the Big Five, trait self-control, and MISCS responses grounds the ratings in the reported conflicts.

Persona regimes. RQ2 asks whether richer information about the person improves the predictions, a persona effect that LLM studies quantify (Hu & Collier, 2024); Tasks 1–3 therefore present each episode under four cumulative persona regimes, which add the source study's baseline measures from the most general to the most self-control-specific. P1 contains demographics and life status, as in common persona prompts (Argyle et al., 2023). P2 adds the Big Five items, a standard description of stable individual differences (Soto & John, 2017) from which LLMs reproduce how people's scores on other scales correlate (Liu et al., 2026). P3 adds the trait self-control items, whose initiation, continuation, and inhibition subscales correspond to the three conflict types (de Boer et al., 2011; Hoyle & Davisson, 2016), and P4 the MISCS items, the person's self-reported metacognition. Each block lists item texts with verbal responses. Because the regimes are cumulative, adjacent contrasts (P2−P1, P3−P2, P4−P3) isolate what each block adds, and P4−P1 gives the total effect. Tasks 1–3 withhold the person's other reports, as persona prompts do, so they test inference from a person description; the earlier-report predictors show what the history adds.

Models. We evaluate 22 LLMs, 10 closed models from the Claude, Gemini, and GPT families and 12 open-weight models from the DeepSeek, Gemma, GLM, Llama, and Qwen families (Appendix Table 22). Each model answers 90,336 prompts without a system prompt at temperature 0 (GPT-5 models at their default temperature), and parsed labels or numbers are scored (Appendices B.2 and B.3). Each prompt is given in German, the participants' language, and in English, from source translations; with reports and scoring fixed, language contrasts test prompt-language sensitivity.

3.3 Evaluation

Metrics. For Tasks 1–2, the Jaccard overlap $J_e = |G_e \cap \hat{G}_e| / |G_e \cup \hat{G}_e|$ between the reported set G_e and the predicted set $\hat{G}_e$ of episode e penalizes missed and added strategies; we also report precision, recall, F1, set size, and label direction, the sign of a label's selection rate minus its report rate. Task 3 uses absolute error and correlates predicted with reported success between people (Spearman ρ of participant means) and within a person (Pearson r of deviations from the person's mean); Task 4 uses Spearman ρ across participants. Episode metrics are first averaged within participants.

Reference predictors. To separate a hard task from a model failure, we compare the LLMs with predictors that never see the focal report (Table 2). Fit on other participants (five participant-grouped folds), a label-frequency rule returns the three most prevalent labels (the rounded mean set size) or as many labels as the model selected, and a logistic ranker returns the top three labels given P1 fields and the episode; for success, a constant returns the mean and a ridge regression predicts it from the episode and reported strategies. The person's own earlier or other reports give their three most frequent labels or mean success, information absent from the prompts, and Task 4 adds single history features and a ridge regression on 31 history features (Appendix A.7).

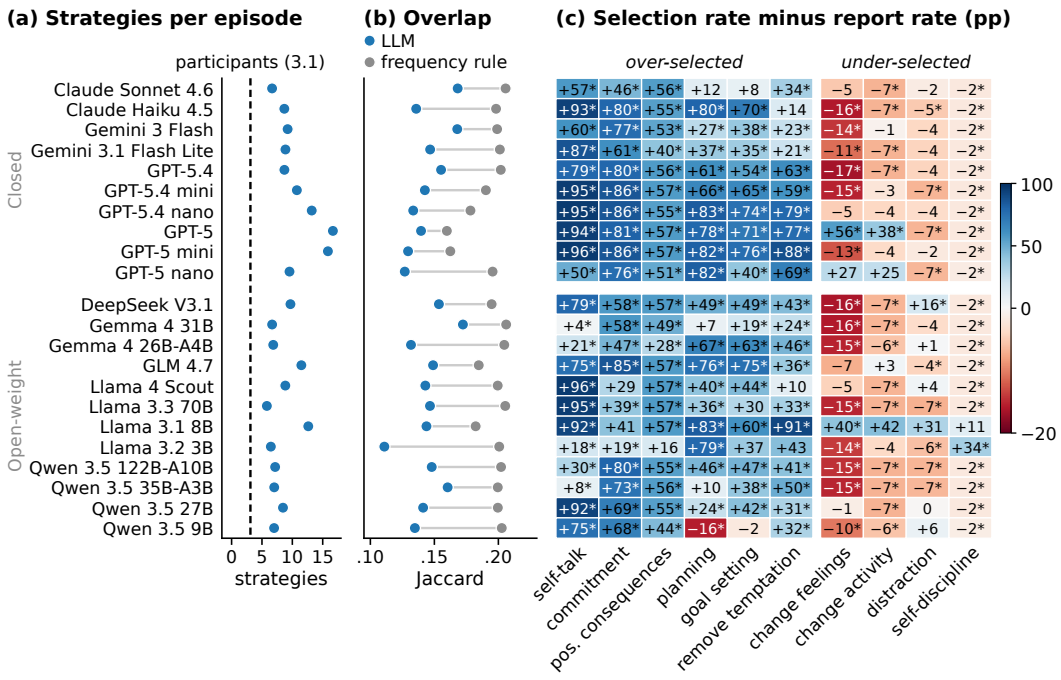

Figure 1: **Every model lists more strategies than participants report, skewed toward internal and planning strategies, and recovers the reported set less well than a label-frequency rule of the same size.** Task 1, persistence, full persona (P4); rows are models (Appendix Table 22); values average English and German. (a) Dashed line: participants' mean. (b) Jaccard with the reported set for the LLM (blue) and the frequency rule cut at the model's set size (grey). (c) Six over- and four under-selected labels with the most Holm-significant conditions; pp = percentage points; $*$ = Holm-significant in both languages, same sign; self-discipline = training outside the activity.

Human evaluators. To test whether the prompts carry recoverable information, nine evaluators (five men, four women; mean age 27.2) answered the English P1 prompts for five Task 1 and five Task 3 persistence episodes of ten participants, stratified by success (Appendix A.2).

Statistics. Persona contrasts use paired Wilcoxon signed-rank tests on participant-level differences, label directions exact McNemar tests on episode-level pairs (checked with participant-level Wilcoxon tests), and Task 4 Spearman tests, each Holm-corrected ($\alpha = .05$; Holm, 1979) within prespecified families, such as the 132 model–language–conflict-type conditions (44 per conflict type) of one persona contrast, task, and metric. Task 4 adds equivalence tests at $\pm.10$ on the Fisher-z scale (TOST; Lakens, 2017), JZS Bayes factors (Rouder et al., 2009), and participant bootstrap intervals for the subscale difference (Appendices A.7, B.5, and I).

4 Results

4.1 Strategy Selection Follows a Generic Repertoire

RQ1 concerns the selection step (action). If the models reproduced how a person selects strategies in a situation, their predicted sets should differ between people and exceed a rule that ignores the person; if they drew on general knowledge alone, they should return similar sets for different people, which need not match report rates. Figure 1 shows the results (example episode: Appendix D).

At this step, the models list more strategies than participants report, and the added labels are unevenly distributed. With P4, every condition predicts more strategies than participants report (Task 1: 5.6–17.0 per episode against 3.1, Figure 1(a); Task 2: 4.3–13.5; 44/44 each). Most added labels are internal and planning strategies, together with removing temptations: self-talk, reported in 4.3% of

Table 2: LLM recovery against reference predictors (persistence; participant-balanced means). Earlier-report rows use the person's earlier reports and the other-episode row all other reports, never the focal one (parentheses: LLM median on the same episodes or range over conditions); human rows compare nine evaluators with the 22 LLMs' English P1 outputs on five episodes per task.

Predictor	Information used	Episodes / participants	Score
Task 1: reported strategy set (Jaccard, higher is better)			
LLMs at P4, median (range)	P4 + episode	1,135 / 353	.142 (.104–.177)
Label frequency, top 3	training-fold prevalence	1,135 / 353	.204
Label frequency, model's set size	training-fold prevalence	1,135 / 353	.199 (.159–.208)
Logistic ranker, top 3	P1 fields + episode	1,135 / 353	.170
Earlier reports, top 3	earlier reports	661 / 224	.257 (LLMs .158)
Humans vs. LLM mean (P1)	same five episodes	5 / 5	.492 vs. .149
Task 3: subjective success, 1–7 (absolute error, lower is better)			
LLMs at P4, median (range)	P4 + episode + strategies	1,130 / 351	1.306 (1.165–1.451)
Earlier-report mean	earlier reports	779 / 245	1.121 (LLMs 1.313)
Other-episode mean	other reports	1,024 / 245	1.061 (LLMs 1.294)
Population mean	other participants	1,130 / 351	1.138
Ridge regression	episode + strategies	1,130 / 351	1.084
Humans vs. LLM mean (P1)	same five episodes	5 / 5	1.500 vs. 1.891

episodes, is selected in a median of 87.5% and over-selected in 44 of 44 conditions, and positive consequences, commitment, removing temptations, goal setting, and planning are over-selected in 39 to 43. Strategies that adjust feelings or the activity in the moment are under-selected despite the larger sets: changing feelings (34 conditions), changing the activity (32), and distraction (29; Figure 1(c)), directions that participant-level tests confirm (33, 32, and 23; Appendix B.4).

The selection also varies little across people: predictions for two different participants overlap at a median $J = .55$ (.59 at P1), against .10 between their reports and .142 between a prediction and the person's own report, and despite their larger size, the 1,135 predicted sets take fewer distinct forms than the reported sets in 42 of 44 conditions (median 312.5 vs. 602; Appendix A.7). All five labels expected before the analysis to be over-selected, on the assumption that internal strategies are frequent in psychology and self-help writing, are over-selected in 25 to 44 conditions; of five expected to be under-selected, only self-discipline training is (41 conditions), which participant-level tests do not confirm (Appendix B.4). When the model lists as many strategies as were reported, self-talk remains over-selected in 32 of 40 conditions (Appendix E). Knowing the reported appraisal shortens the lists and leaves recovery unchanged: in Task 2, the models list fewer strategies (median 6.5 vs. 8.7) and recover the reported set at the same level (median Jaccard .143 vs. .142).

Every model recovers the reported set less well than the frequency rule (Figure 1(b), Table 2). The median P4 condition recalls about half of the reported strategies (.512) at precision .178 (Jaccard .142, F1 .229); the top-3 frequency rule recalls .397 but has higher precision (.330), Jaccard (.204), and F1 (.300) than all 44 conditions. The longer lists do not explain this deficit: truncated at each model's own set size, the frequency ranking has higher Jaccard (median .199), precision (.234), and recall (.673) than the model in all 44 conditions (Figure 1(b)). A logistic ranker with P1 fields and the episode context reaches Jaccard .170, and the person's earlier reports reach .257 on the 661 episodes where they exist, on which the LLMs reach at most .197. In each family with several evaluated models, the largest or highest-tier model recovers more than the smallest (Appendix G).

Human evaluators recover more from the same P1 prompts: on five episodes, nine evaluators reached a mean $J = .492$ (identical sets in 33.3% of judgments), the 22 LLMs .149, the best model .220, and the frequency rule .097, so the P1 prompt carries recoverable information on these episodes. At the selection step, the models therefore return a generic repertoire weighted toward internal and planning strategies, which misses the person's situation-specific choice, matches the reported set poorly, and departs even from the population's report rates. Because the frequency rule and the human evaluators recover more from the same or less information, the gap lies in the models' selection step; Section 4.2 tests whether this pattern depends on the conflict type, the prompt, or the cohort.

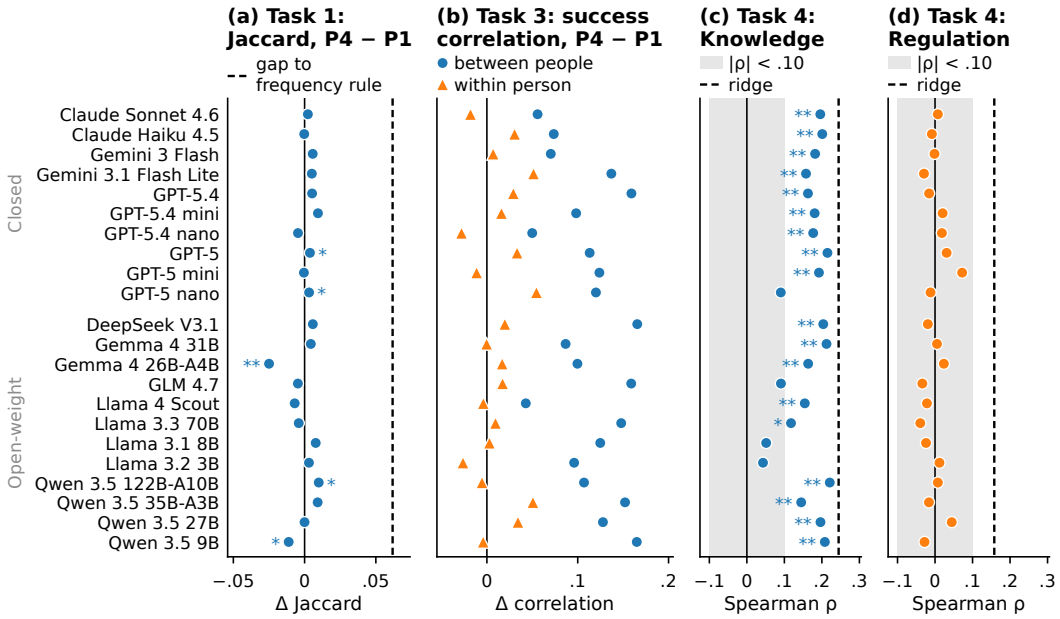

Figure 2: **Person information improves appraisal between people and changes selection and within-person appraisal only slightly; metacognition ratings from histories rank Knowledge weakly and miss Regulation.** Persistence; rows and means as in Figure 1 (values in Table 41); ** and * mark Holm significance (a: P4 vs. P1; c, d: $\rho \neq 0$) in both languages or in one. (a) Task 1 Jaccard, P4 minus P1; dashed: the gain the median P4 model needs to reach the frequency rule. (b) Task 3: P4 minus P1 change in the between-person (participant means) and within-person (deviations from the mean) correlation of predicted with reported success. (c, d) Task 4: Spearman ρ with observed Knowledge (c) and Regulation (d); grey band: $|\rho| < .10$; dashed: ridge regression.

4.2 The Selection Pattern Recurs Across Settings

A repertoire that the models bring to any conflict should produce the same label directions across conflict types, prompts, and cohorts. In initiation and inhibition, every P4 condition lists more strategies than participants report (5.6–17.4 vs. 3.3 and 4.5–15.8 vs. 2.8 strategies), with self-talk, positive consequences, and commitment over-selected in 43–44 of 44 conditions (Appendix H).

Prompt settings change individual outputs, and the pattern persists. In six models rerun at P4 on 269 episodes from 88 participants, temperature $T = 0.7$ and shuffled options left Task 1 Jaccard near the original .168 (.166 and .162), every model listed more strategies than reported, and the label direction stayed the same for 96.5% and 85.1% of model-label pairs (Appendix A.4). English and German prompts give similar Jaccard (.148 and .141) and label directions (84% of model-label pairs agree; Appendix J).

In the U.S. food-desire data (nine labels), four models predicted 3.0–4.7 strategies against 1.7 reported ($J = .217$–$.247$) and chose internal strategies, removing the food, and distraction more often, and suppression less often, than participants, matching the German directions for internal strategies, removing temptations, and suppression (Appendix A.6). The larger sets and the skew toward internal strategies thus recur across conflict types, prompts, languages, and cohorts; Section 4.3 asks whether person information changes them.

4.3 Person Information Improves Appraisal Between People

RQ2 asks whether person information enters the process steps. If it entered the selection step, the added blocks, in particular trait self-control and the MISCS, which describe how a person regulates, should raise Task 1 recovery. They change it only slightly (Figure 2(a)): the median P4−P1 Jac-

card change is $+.001$ (favorable in 25 of 44 conditions, three Holm-significant in each direction; Appendix F), and in initiation and inhibition the median change stays below .01 (Appendix H).

At the appraisal step, person information lowers the error of predicted success: from P1 to P4, it falls in 35 of 44 persistence conditions (median .055, rank-biserial r .10; nine Holm-significant, none unfavorable; Appendix K), mostly when the Big Five and trait self-control blocks are added (favorable in 35 and 31 of 44 conditions; MISCS block, 13 of 44). The improvement lies mostly between people (Figure 2(b)): the correlation of a person's mean predicted with mean reported success rises in all 44 conditions, from a median $\rho = .14$ at P1 to .26 at P4, whereas the within-person correlation stays at $r = .08$–.10. A ridge regression on the episode and reported strategies reaches .31 and .24, so each episode's strategies carry between- and within-person information about its appraisal that the models track weakly. In inhibition and initiation, error rises (37 and 26 conditions), and in inhibition the between-person correlation falls from .14 to .11 (Appendices A.7 and H).

In absolute error, predicted success remains less accurate than simple references (Table 2): other participants' mean success (1.138) and the ridge regression (1.084) err less than every P4 condition (1.165–1.451), and the person's earlier mean (1.121) and other-episode mean (1.061) err less than all 44 conditions on the same episodes. Removing the persona or the episode (four models) raised the pooled error from 1.31 to 1.41 and 1.72, and adding the person's mean lowered it to 1.19 (Appendix A.3); the models therefore use the prompt content, while shuffled strategy order and $T = 0.7$ left the error stable (Appendix A.4). In persistence conflicts, person information thus enters the appraisal step as a stable success level for each person, while selection and the appraisal of each episode change only slightly. The conflict history records how the person selected and appraised strategies across conflicts; Section 4.4 tests whether metacognition ratings reflect it.

4.4 Metacognition Ratings Follow Outcomes

RQ3 concerns the metacognitive step, which spans conflicts. If the models' ratings reflected how the person plans, monitors, and evaluates strategy use, they should rank people by MISCS Regulation, whose items describe choosing strategies, analyzing their usefulness, and monitoring progress. They do not (Figure 2(c, d); 466–467 participants, median seven conflicts each). The median Knowledge correlation is $\rho = .178$, positive in all 44 conditions and Holm-significant in 35; the median Regulation correlation is .003, no condition is Holm-significant, all 44 satisfy $|\rho| < .10$, TOST supports equivalence within $\pm.10$ in 25, Bayes factors favor a zero association in 43, and the Knowledge–Regulation difference has a bootstrap interval above zero in 36 (Appendices A.7 and I).

Differences in reliability do not explain the contrast. Regulation is at least as internally consistent as Knowledge (α .79 vs. .72; Appendix A.5), correcting for the unreliability of the observed scores moves the Knowledge median to about .21 and the Regulation equivalence bound to about .11, and the contrast persists under shuffled order and $T = 0.7$ (Appendix A.4).

Simple references show which parts of the history the ratings track. A person's mean reported success ranks Knowledge at $\rho = .234$ alone, above 43 of 44 conditions, but Regulation at $-.007$, and the models' Knowledge and Regulation predictions correlate with this mean (median $\rho = .45$ and .34) and with strategies per episode (.46 and .52); controlling for these two features and repertoire size (distinct strategies reported) removes the Knowledge association (median partial $\rho = -.02$). A ridge regression fit to other participants ranks Knowledge at .245, near the best LLM condition (.235), and Regulation at .158, above every LLM condition (maximum .096), so the histories carry Regulation information (Appendix A.7); of the single features, only repertoire size relates to Regulation (.09). These summaries thus account for the ratings' Knowledge association, and the ratings miss the process information from which the ridge ranks Regulation.

5 Discussion

Across the three steps, the models capture what people generally do and how people differ, and miss how, for a specific person, the situation shapes strategy choice, strategies shape appraisal, and strategy use reflects metacognition; each step can look plausible to outcome-based evaluations while this process is missing or weak.

Selection follows a repertoire. People match strategies to who they are and what the situation demands (Mischel & Shoda, 1995; Wenzel et al., 2024), whereas the models return the deliberate strategies expected before the analysis to be frequent in psychological and self-help writing, which change little with either and recover the reported set less well than a frequency rule. Prior studies rated LLM strategy advice as plausible or helpful (Zhan et al., 2024; Jörke et al., 2025); at the person–episode level, this knowledge is generic: it names which strategies exist and leaves out when a person uses them, in line with LLMs' expectation that people choose more rationally than they do (Liu et al., 2025). Developers of LLM coaches should therefore also test whether recommendations change with the person and the situation, since plausibility ratings would pass a coach that gives every user the same strategies; researchers who simulate participants with LLMs would obtain strategy data with fewer individual differences (312.5 vs. 602 distinct sets), extending the flattening of identity groups (Wang et al., 2025) to individuals; and person–episode behavioral records (Binz et al., 2025) could teach models how situations shape a person's strategy choice.

Appraisal follows the person. In daily life, how well an episode goes depends on what a person does in that situation (Wenzel et al., 2024). The models' appraisals track the person's typical success, which person information improves, and barely the episode's strategies (within-person $r = .08$–$.10$ against .24 for a ridge regression), and knowing the outcome leaves strategy recovery unchanged. Persona studies have asked whether person information improves predictions (Hu & Collier, 2024); the step-wise test shows where it goes: into a prior on how successful the person is, while the relation between an episode's strategies and its appraisal stays weak in both directions. For adaptive interventions, which must decide when to offer support, such predictions indicate who tends to struggle and carry little signal about when; researchers who simulate experience-sampling or diary data with LLMs would obtain attenuated within-person effects, such as the relation of strategy use to success, even where between-person structure is reproduced (Liu et al., 2026).

Metacognition follows outcomes. Metacognition rests on how a person selects, monitors, and evaluates strategies across conflicts (Hennecke & Bürgler, 2023). The models' ratings of both subscales follow success and strategy counts: their Knowledge correlation ($\rho = .178$) vanishes once these summaries and repertoire size are controlled, and Regulation, which success does not reveal and the history carries (ridge $\rho = .158$), is missed. LLM inferences of dispositions have been validated by correlation with questionnaire scores (Peters & Matz, 2024); the Knowledge ratings would pass this test, so correlation-based validation can certify an inference from outcomes as an inference about metacognition. Developers of LLM-based assessment in education or clinical screening should therefore also test whether ratings hold beyond outcome summaries, since outcome-based ratings would score a struggling, well-regulated person as low; for self-control research, regulation is the facet that outcomes hide, which process measures such as strategy variety or momentary planning and monitoring reports (Bürgler & Hennecke, 2024) could capture.

Limitations. The four-task test covers one German cohort, one strategy list, and one metacognition scale; it needs only multi-label action reports, an appraisal, and person measures and could be applied to other cultures and to experience-sampling studies of emotion regulation (Heiy & Cheavens, 2014; Blanke et al., 2020). The design locates the failures at the reported steps and leaves their sources open: base and instruction-tuned checkpoints of one model family could separate pretraining from tuning, and the robustness checks, which cover four to six models, could be extended to models of known size and to models fine-tuned on experience-sampling records.

6 Conclusion

This study tested whether LLMs reproduce the process by which a person handles everyday self-control conflicts, scoring each step against the person's own experience-sampling reports. Given person descriptions, the 22 LLMs predicted a generic strategy repertoire and ranked people by typical success while weakly tracking the variation across a person's episodes; given conflict histories, they rated metacognition from success while missing regulation. The models thus capture what people generally do and how people differ, and miss how, for a specific person, the situation shapes strategy choice, strategies shape appraisal, and strategy use reflects metacognition, failures that outcome-based evaluations can miss. Testing this process directly could guide the use of LLMs as models of individuals in cognitive science and self-regulation support.

AI USE STATEMENT

The LLMs evaluated in this study are its objects of study and were used only for inference, to generate the predictions analyzed here. During writing, we used generative AI tools to correct grammar and find typos. We reviewed all AI-assisted edits and take responsibility for the final content of this paper, including text, claims, and artifacts.

ETHICS STATEMENT

This study analyzes pseudonymized experience-sampling data from a published study (Bürgler & Hennecke, 2024; Wenzel et al., 2024), accessed through the study's public OSF project (`https://osf.io/q5hzg`); the published article and OSF preprint associated with the dataset are licensed under CC BY 4.0. Participants were not recontacted.

REPRODUCIBILITY STATEMENT

The code for prompt generation, inference, scoring, and all reported analyses is provided as supplementary material. Section 3 and Appendix B specify the data fields, task schemas, all four prompt templates (Appendix B.2), the model list, decoding settings, and output parsing (Appendix B.3), metric and missing-data rules (Appendix B.4), and the statistical families (Appendices B.5 and C). Appendix A documents the reference predictors, the human-evaluator protocol, the order, temperature, and input-removal checks, the external replication, and the size-matched, success-tracking, and Task 4 reference analyses. The source data are available from the source study's public OSF project.

REFERENCES

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025. doi: 10.1038/s41586-025-09215-4.

Elisabeth S. Blanke, Annette Brose, Elise K. Kalokerinos, Yasemin Erbas, Michaela Riediger, and Peter Kuppens. Mix it to fix it: Emotion regulation variability in daily life. *Emotion*, 20(3): 473–485, 2020. doi: 10.1037/emo0000566.

Sebastian Bürgler and Marie Hennecke. Metacognition and polyregulation in daily self-control conflicts. *Scandinavian Journal of Psychology*, 65(2):179–194, 2024. doi: 10.1111/sjop.12964.

Sebastian Bürgler, Rick H. Hoyle, and Marie Hennecke. Flexibility in using self-regulatory strategies to manage self-control conflicts: The role of metacognitive knowledge, strategy repertoire, and feedback monitoring. *European Journal of Personality*, 35(6):861–880, 2021. doi: 10.1177/0890207021992907.

Sebastian Bürgler, Kristian Kleinke, and Marie Hennecke. The Metacognition in Self-Control Scale (MISCS). *Personality and Individual Differences*, 199:111841, 2022. doi: 10.1016/j.paid.2022.111841.

Benjamin J. de Boer, Edwin A. J. van Hooft, and Arnold B. Bakker. Stop and start control: A distinction within self-control. *European Journal of Personality*, 25(5):349–362, 2011. doi: 10.1002/per.796.

Jane E. Heiy and Jennifer S. Cheavens. Back to basics: A naturalistic assessment of the experience and regulation of emotion. *Emotion*, 14(5):878–891, 2014. doi: 10.1037/a0037231.

Marie Hennecke and Sebastian Bürgler. Metacognition and self-control: An integrative framework. *Psychological Review*, 130(5):1262–1288, 2023. doi: 10.1037/rev0000406.

Marie Hennecke, Thomas Czikmantori, and Veronika Brandstätter. Doing despite disliking: Self-regulatory strategies in everyday aversive activities. *European Journal of Personality*, 33(1): 104–128, 2019. doi: 10.1002/per.2182.

Wilhelm Hofmann, Roy F. Baumeister, Georg Förster, and Kathleen D. Vohs. Everyday temptations: An experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology*, 102(6):1318–1335, 2012. doi: 10.1037/a0026545.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Rick H. Hoyle and Erin K. Davisson. Varieties of self-control and their personality correlates. In Kathleen D. Vohs and Roy F. Baumeister (eds.), *Handbook of Self-Regulation: Research, Theory, and Applications*, pp. 396–413. Guilford Press, New York, third edition, 2016.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, 2024. doi: 10.18653/v1/2024.acl-long.554.

Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. *The Big Five Inventory–Versions 4a and 54*. University of California, Berkeley, Institute of Personality and Social Research, 1991.

Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. GPTCoach: Towards LLM-based physical activity coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–46, 2025. doi: 10.1145/3706598.3713819.

Dave S. Kerby. The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3:11.IT.3.1, 2014. doi: 10.2466/11.IT.3.1.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024. doi: 10.1073/pnas.2405460121.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Ryan Liu, Jiayi Geng, Joshua C. Peterson, Ilia Sucholutsky, and Thomas L. Griffiths. Large language models assume people are more rational than we really are. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=dAeET8gxqg`.

Yi-Fei Liu, Yi-Long Lu, Di He, and Hang Zhang. From five dimensions to many: Large language models as precise and interpretable psychological profilers. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=JXFnCpXcnY`.

Richard B. Lopez, Danielle Cosme, Kaitlyn M. Werner, Blair Saunders, and Wilhelm Hofmann. Associations between use of self-regulatory strategies and daily eating patterns: An experience sampling study in college-aged women. *Motivation and Emotion*, 45(6):747–758, 2021. doi: 10.1007/s11031-021-09903-4.

Marina Milyavskaya, Blair Saunders, and Michael Inzlicht. Self-control in daily life: Prevalence and effectiveness of diverse self-control strategies. *Journal of Personality*, 89(4):634–651, 2021. doi: 10.1111/jopy.12604.

Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2):246–268, 1995. doi: 10.1037/0033-295X.102.2.246.

Heinrich Peters and Sandra C. Matz. Large language models can infer psychological dispositions of social media users. *PNAS Nexus*, 3(6):pgae231, 2024. doi: 10.1093/pnasnexus/pgae231.

Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, 2009. doi: 10.3758/PBR.16.2.225.

Katja Schlegel, Nils R. Sommer, and Marcello Mortillaro. Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology*, 3:80, 2025. doi: 10.1038/s44271-025-00258-x.

Gregory Schraw and Rayne Sperling Dennison. Assessing metacognitive awareness. *Contemporary Educational Psychology*, 19(4):460–475, 1994. doi: 10.1006/ceps.1994.1033.

Christopher J. Soto and Oliver P. John. Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, 68:69–81, 2017. doi: 10.1016/j.jrp.2017.02.004.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. doi: 10.1038/s41562-024-01882-z.

June P. Tangney, Roy F. Baumeister, and Angie Luzio Boone. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, 72 (2):271–324, 2004. doi: 10.1111/j.0022-3506.2004.00263.x.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309, 2023. doi: 10.1111/cogs.13309.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025. doi: 10.1038/s42256-025-00986-z.

Mario Wenzel, Sebastian Bürgler, Veronika Brandstätter, Antonia Kreibich, and Marie Hennecke. Self-regulatory strategy use, efficacy, and strategy-situation-fit in self-control conflicts of initiation, persistence, and inhibition. *European Journal of Personality*, 38(2):189–208, 2024. doi: 10.1177/08902070221150478.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond C. Ong. Large language models are capable of offering cognitive reappraisal, if guided. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=yK8MT91dQY.

APPENDIX OVERVIEW

Table 3 maps each main-text claim to the appendix that supports it.

Table 3: Map from main-text results to supporting appendices.

Main-text result	Appendix
Reference predictors and human evaluators (Table 2)	A.1, A.2, A.3
Order, temperature, input removal, and external replication (Sections 4.2 and 4.3)	A.4, A.3, A.6
Size-matched frequency rule, cross-person overlap, success tracking, Task 4 references (Sections 4.1–4.4)	A.7
MISCS reliability (Section 4.4)	A.5
Data, tasks, prompts, models, metrics, statistics (Section 3)	B
Units of analysis and correction families (Section 3.3)	C
Example episode and strategy groups (Sections 3.1 and 4.1)	D
All labels, set size, matched-cardinality slice (Section 4.1)	E
Persona-block contrasts (Section 4.3)	F
Model-level and family results (Sections 4.1 and 4.3)	G
Initiation and inhibition (Sections 4.2 and 4.3)	H
Task 4 construct, equivalence, and coverage audits (Section 4.4)	I
Subgroup, model-family, and language audits (Section 4.2)	J
Effect sizes and equivalence tests (Sections 4.3 and 4.4)	K

A REFERENCE PREDICTORS AND ROBUSTNESS CHECKS

This appendix documents the reference predictors, the human-evaluator comparison, and the robustness checks summarized in Sections 3.3–4.4. All values are participant-balanced unless noted.

A.1 REFERENCE PREDICTORS FOR TASK 1

These predictors measure how much of the reported strategy set can be recovered without the focal report (Table 4). The frequency and logistic predictors use five participant-grouped folds, so all episodes of a participant are either in training or in test; $k = 3$ is the rounded mean reported set size of the training folds. The logistic ranker is a one-vs-rest logistic regression per label on age, sex, marital status, education, current status, conflict strength, activity, and goal indicators. Each per-label model is an L2-penalized logistic regression with $C = 1$ (liblinear solver); within each training fold, age, conflict strength, and the current-status, activity, and goal indicators are divided by their standard deviation; marital status, education, and sex are one-hot encoded; missing age (15 of 1,135 episodes) and marital status (3) take the training-fold median and mode; a missing goal domain (138) forms its own indicator; and a label with one class in the training fold receives that class as a constant score. The folds are shuffled with seed 240725. Earlier-report predictors are evaluated on episodes for which an earlier report is available and predict each episode from the same participant's earlier persistence reports (all other reports in the diagnostic row), which lets them draw on the person's own history. The first two rows are size-matched diagnostics: both take the focal set's true size from the report as an oracle input, which fixes set size and gives chance and frequency levels for label choice at that size.

The held-out frequency predictor reaches a participant-balanced F1 of .300 and the logistic ranker .253.

Table 4: Task 1 reference predictors (persistence). "Oracle size" predictors use the size of the focal reported set. Rows with 994 episodes use the success-scored reports that list at least one strategy. The pair row for reports with strategies uses the 496,453 different-participant pairs among the 999 persistence reports that list at least one strategy. PB: participant-balanced mean; EW: episode-weighted mean.

Predictor	Episodes	Participants	Jaccard (EW)	Jaccard (PB)
Random set of the reported size (oracle size)	994	333	.092	–
In-sample frequency, top $\|G_e\|$ (oracle size)	994	333	.272	.273
Held-out frequency, top 3	1,135	353	.198	.204
Held-out logistic ranker (P1 fields), top 3	1,135	353	.167	.170
Leave-one-participant-out frequency, top 3	994	333	.225	.231
Held-out logistic ranker, top 3 (994-episode set)	994	333	.189	.189
Earlier reports (strictly earlier), top 3	661	224	.261	.257
Other reports of the same participant, top 3 (diagnostic)	885	224	.273	.263
LLMs at P4, median over 44 conditions (range)	1,135	353	–	.142 (.104–.177)
Different-participant pairs, reports with strategies, mean (median)	–	–	.128 (.083)	–
Different-participant pairs, at least one report with strategies	–	–	.100	–

A.2 Human Evaluators

This comparison measures how well people recover the reported strategies and success from the information the LLMs receive. Nine human evaluators (five men and four women; mean age 27.2 years, SD 3.9) answered the same English P1 prompts as the LLMs (Figures 3 and 5): age, sex, marital status, education, current status, the activity being continued, its goal domain, and conflict strength. Personality, trait self-control, and MISCS items were not shown. Ten persistence episodes from ten different participants were selected by deterministic stratified sampling (seed 20260727) among episodes with a success score, at least one reported strategy, a recorded activity, and at least three scored persistence episodes of the same participant, stratified by success (low 1.0–3.0: three; middle 3.5–5.0: four; high 5.5–7.0: three) and diversified by activity, goal domain, and reported set size. Five episodes were assigned to Task 1 and five different episodes to Task 3, and every evaluator answered all ten in the same order.

For Task 1, evaluators selected all strategies that apply from the same 25-option list as the LLMs; for Task 3, they saw the reported strategies and gave a single 1–7 success rating, as the LLMs did, which is scored against the mean of the two source items. Exact-set match equals 1 only when the selected set is identical to the reported set.

Across the 45 Task 1 judgments (Table 5), the exact-set match rate was .333. Across the 45 Task 3 predictions, the root-mean-square error was 2.159. For the matched comparison, the English P1 outputs of the 22 LLMs on the same episodes (110 predictions per task) reached Jaccard .149 and F1 .233 on Task 1 (best model .220) and absolute error 1.891 on Task 3 (model range 1.700–2.100). On the same five Task 1 episodes, the top-3 frequency rule reaches Jaccard .097; two of the five episodes report a single strategy (changing the activity).

A.3 Task 3 Earlier-Report Predictors With Input Isolation

These predictors test whether Task 3 accuracy reflects the focal episode or the person's general rating level (Table 6).

Table 5: Human-evaluator results by episode (nine evaluators per episode). Task 3 history is the source participant's mean success over other persistence episodes.

Task 1 (reported strategy set)				
Episode	Activity	Reported strategies	Jaccard	F1
E01	university course	6	.605	.687
E02	family or friends	1	.421	.463
E04	housework	3	.496	.563
E06	eating or drinking	7	.439	.503
E09	sports	1	.501	.539
All 45 judgments			.492	.551
Task 3 (subjective success, 1–7)				
Episode	Report	Human mean prediction	Human MAE	History error
E03	2.0	4.000	2.000	2.000
E05	3.5	3.889	1.056	0.643
E07	4.5	4.222	0.722	0.500
E08	7.0	4.889	2.111	3.625
E10	5.5	4.222	1.611	1.000
All 45 predictions			1.500	1.554

Table 6: Task 3 reference predictors (persistence). History predictors use the same participant's other or earlier success reports, which the LLM prompts withhold, so they show what the person's history adds. The two LLM rows are evaluated on the same episodes as the other-episode (1,024) and earlier-episode (779) predictors, respectively; their ranges over the 44 conditions are 1.151–1.467 (1,024 episodes) and 1.162–1.484 (779 episodes).

Predictor	Episodes	Participants	MAE (EW)	MAE (PB)	Spearman ρ
Mean of other participants' reports	1,130	351	1.136	1.138	–
Participant's other episodes (focal excluded)	1,024	245	0.977	1.061	.481
Participant's strictly earlier episodes	779	245	1.031	1.121	.439
LLMs at P4, median of 44	1,024	245	–	1.294	–
LLMs at P4, median of 44	779	245	–	1.313	–

None of the 44 P4 conditions has a lower participant-balanced error than the corresponding history predictor on either coverage.

The input-isolation run used Claude Haiku 4.5, GPT-5.4 mini, GPT-5.4 nano, and GPT-5 mini under P4 in English and German on 49 randomly selected participants with three or four scored persistence episodes (165 episodes, 164 scorable; seed 240725). The persona-only condition asks for one expected success score from the persona card and applies it to all of that participant's episodes; the episode-only condition removes the persona card; the full condition is the Task 3 prompt; and the full-plus-mean condition adds the mean of the participant's other scored episodes, excluding the focal one (Table 7).

Table 7: Task 3 input isolation. Pooled rows average over all numeric outputs; Claude Haiku 4.5 returned no number for the persona-only prompts of 146 of 328 episode rows (English 119/164, German 27/164; each GPT model 0/328), and its model row reports the three inputs that contain the episode.

	Persona only	Episode only	Full	Full + mean
Pooled MAE	1.722	1.408	1.312	1.187
Pooled Spearman ρ	.049	.082	.113	.268
Pooled signed bias	-1.139	-0.358	-0.354	-0.198
Claude Haiku 4.5	–	1.409	1.323	1.195
GPT-5.4 mini	1.482	1.317	1.223	1.101
GPT-5.4 nano	1.759	1.497	1.399	1.320
GPT-5 mini	1.503	1.415	1.302	1.131

For each GPT model, the full input has lower error than either isolated input, and for all four models the participant mean lowers error further.

A.4 Robustness Checks

These checks test whether the label-selection pattern depends on one option order or on deterministic decoding. Six models (Claude Sonnet 4.6, Gemini 3 Flash, GPT-5.4, Gemma 4 31B, Llama 3.3 70B, Qwen 3.5 35B-A3B) answered Tasks 1–2 under P4 in English and German for 269 persistence episodes from 88 participants (a participant-level 25% sample, seed 240725). The original condition used the main-run option order at $T = 0$; three conditions at $T = 0$ shuffled the entire 25-option list independently per episode; five repetitions at $T = 0.7$ kept the original order. The $T = 0.7$ runs set top-p to 1.0 for every model except Claude Sonnet 4.6, whose requests set temperature only, and the $T = 0$ runs kept the settings of Table 23. The Tasks 1–2 reruns added 51,648 responses. Tasks 3–4 list the reported strategies in the prompt, so five models (the same panel without Llama) answered them with that list shuffled (three conditions) or at $T = 0.7$ (five repetitions), for 279 Task 3 episodes (all persistence episodes of the 88 participants; Tasks 1–2 use the 269 whose original outputs are matched to the reruns, which pairs each perturbed output with its original) and 88 Task 4 histories (22,320 added Task 3 responses and 7,040 added Task 4 completions). These reruns treat an empty Task 3 response as missing and clip Task 4 item answers to 1–5. Tables 8 and 9 report Tasks 1–2, and Table 10 reports Tasks 3–4.

Pooled Task 3 signed bias was -0.027 (original), -0.018 ($T = 0.7$), and -0.025 (shuffled).

A.5 MISCS Subscale Reliability

This analysis tests whether lower measurement reliability accounts for the Knowledge–Regulation contrast (Table 11). The values below match those published for this sample (Knowledge $\alpha = .72$, Regulation $\alpha = .79$; Bürgler et al., 2022).

The evaluation file contains one MISCS administration. In the source study, the subscales correlate .54 (Knowledge) and .51 (Regulation) across two weeks (Bürgler et al., 2022), so one administration also carries occasion-specific variance.

Table 8: Tasks 1–2 under perturbation, pooled over six models and two languages. Participants reported 3.239 strategies per episode in this sample. Agreement is the Jaccard overlap between a perturbed output and the original output for the same prompt; label-profile ρ correlates the per-label selection rates; Retained is the share of model-by-label over- or under-selection signs that stay the same.

Task	Condition	Jaccard	F1	Predicted size	Agreement	Profile ρ	Retained
1	Original	.168	.262	7.211	–	–	–
1	$T = 0.7$	.166	.259	7.086	.805	.985	.965
1	Shuffled options	.162	.250	6.778	.583	.815	.851
2	Original	.166	.258	5.885	–	–	–
2	$T = 0.7$	.165	.256	5.767	.783	.983	.957
2	Shuffled options	.162	.250	5.720	.574	.840	.842

Table 9: Tasks 1–2 under perturbation by model (Jaccard / predicted set size, averaged over languages and repetitions).

Model	T1 original	T1 $T = 0.7$	T1 shuffled	T2 original	T2 $T = 0.7$	T2 shuffled
Claude Sonnet 4.6	.182/6.54	.182/6.60	.154/4.07	.183/4.29	.183/4.33	.166/4.32
Gemini 3 Flash	.175/8.97	.173/8.73	.169/8.83	.174/7.60	.179/7.33	.171/7.34
GPT-5.4	.162/8.76	.161/8.78	.170/7.86	.162/7.32	.162/7.38	.171/6.43
Gemma 4 31B	.176/6.38	.174/6.36	.178/6.24	.183/4.59	.177/4.50	.163/4.43
Llama 3.3 70B	.150/5.66	.155/5.56	.138/6.62	.134/4.88	.139/4.82	.139/5.14
Qwen 3.5 35B-A3B	.164/6.96	.151/6.48	.163/7.05	.161/6.63	.154/6.25	.161/6.66

Table 10: Tasks 3–4 under perturbation by model. Task 3 cells are MAE / Spearman ρ; Task 4 cells are Spearman ρ for Knowledge (K) and Regulation (R).

	Task 3: MAE / ρ			Task 4: K ρ / R ρ		
Model	Original	$T = 0.7$	Shuffled	Original	$T = 0.7$	Shuffled
Claude Sonnet 4.6	1.315/.278	1.318/.265	1.340/.259	.150/.032	.120/.024	.101/.008
Gemini 3 Flash	1.415/.267	1.480/.258	1.463/.253	.183/.007	.135/.004	.183/.076
GPT-5.4	1.276/.301	1.283/.292	1.266/.301	.267/.037	.273/.074	.257/.084
Gemma 4 31B	1.451/.249	1.454/.237	1.453/.247	.199/.050	.173/.013	.188/−.003
Qwen 3.5 35B-A3B	1.507/.154	1.551/.151	1.482/.190	.178/.046	.159/−.026	.257/.082
Pooled	1.393/.250	1.417/.241	1.401/.250	.195/.034	.172/.018	.197/.049

Table 11: Reliability and rank recovery of the two MISCS subscales. Reliability is computed across participants on the six items whose mean is the scored subscale (n = 464 and 463); ω is computed from a one-factor maximum-likelihood factor analysis of the standardized items. The attenuation correction divides by the square root of the outcome reliability only.

Diagnostic	Knowledge	Regulation
Cronbach's α	.719	.786
McDonald's ω	.721	.786
Median Spearman ρ over 44 conditions	.178	.003
Holm-significant conditions	35/44	0/44
Conditions with $\lvert\rho\rvert < .10$	7/44	44/44
TOST equivalence within $\pm.10$	1/44	25/44
Correction factor $1/\sqrt{\alpha}$	1.18	1.13
Corrected median (Knowledge) or corrected band edge (Regulation)	.210	.113

A.6 External Replication of Task 1

This replication tests whether the action-level failure appears outside the German cohort, domain, and label set. The public experience-sampling study of Lopez et al. (2021) contains 1,274 food-desire reports from 106 college-aged women in the United States, with multi-label strategy responses in the Task 1 format, which suits this test. Reports with a present desire and at least some resistance recorded a multi-label strategy response; this filter yields 686 episodes from 102 participants. The prompt contained age, sex, body-mass index, eating-restraint status, desire strength, hunger, goal conflict, and goal importance; resistance intensity, enactment, and amount eaten were withheld because they describe the response or its aftermath. The nine targets are leaving the situation, removing the food, promising another person, monitoring, reminding oneself of goals, accepting the desire, distraction, suppression, and another strategy. Four API models answered in English at $T = 0$ (GPT-5 mini and GPT-5 nano at their default temperature) with the source study's option order; all 2,744 outputs parsed (Table 12).

Table 12: External Task 1 replication on the U.S. food-desire data (participant-balanced). Participants reported 1.738 strategies per episode.

Model	Jaccard	F1	Predicted set size	Size error
Claude Haiku 4.5	.217	.307	3.108	+1.370
GPT-5.4 nano	.227	.342	4.737	+2.999
GPT-5 mini	.247	.347	2.995	+1.257
GPT-5 nano	.218	.313	3.729	+1.991

Participants reported situational or external strategies in .170 of episodes, cognitive or internal strategies in .583, distraction in .456, suppression in .284, and other strategies in .103 (participant-balanced rates). Across the 686 episodes (episode-weighted rates), the four models selected situational or external strategies in 72–100% of episodes (participants 18%), cognitive or internal strategies in 92–100% (56%), distraction in 90–100% (48%), suppression in none (34%), and other strategies in 0–5% (8%). The models therefore selected internal strategies, situational strategies such as removing the food, and distraction more often than participants did, and suppression and other strategies less often.

A.7 Further Reference Analyses

These analyses test the main comparisons against three alternative explanations: set size (Task 1), person-level versus episode-level accuracy (Task 3), and the information in the history itself (Task 4). All analyses reuse stored model outputs and source data; no model was rerun (`scripts/added_checks.py` and `scripts/generic_overlap.py`).

Task 1 at equal set size. For each model–language condition and episode, the frequency rule returns the m most prevalent labels in the training folds of the episode's participant, where m is the number of labels the model selected for that episode; folds and prevalence ranking are those of the top-3 rule (Table 4). At equal set size, the frequency ranking recovers the reported set better than every condition on every metric (Table 13; per-model values in Table 40).

Table 13: Task 1 (persistence, P4): LLMs versus the held-out frequency ranking truncated at each model's own set size (medians over the 44 model–language conditions). The size-matched rule is higher than the LLM on all four metrics in 44 of 44 conditions. Recall is computed over reports that list at least one strategy, for all rows; precision counts an empty predicted set as 0.

Predictor	Set size	Jaccard	Precision	Recall	F1
LLMs	8.69	.142	.178	.512	.229
Frequency, model's set size	8.69	.199	.234	.673	.302
Frequency, top 3	3.00	.204	.330	.397	.300

Table 14: Task 3 success tracking by conflict type (medians over 44 conditions). Between people: Spearman ρ between participants' mean predicted and mean reported success. Within person: one Pearson r over all scored episodes of the conflict type after subtracting each participant's mean prediction and mean report, so each episode carries equal weight, a participant with a single scored episode contributes zero deviations, and a participant with constant reports or predictions contributes only to the other variable's sum of squares (persistence: 1,130 episodes, 1,024 of them from the 245 participants with at least two). "Higher" counts conditions in which P4 exceeds P1. The ridge regression uses the episode context and reported strategies (persistence; participant-grouped folds; absolute error 1.084).

	Between people			Within person		
Conflict type	P1	P4	Higher	P1	P4	Higher
Persistence	.143	.263	44	.082	.097	28
Initiation	.147	.239	43	.119	.154	30
Inhibition	.139	.112	13	.143	.165	27
Ridge regression (persistence)	.306			.243		

Table 15: Task 4 references (Spearman ρ with observed MISCS scores; 467 participants, 466–467 in LLM conditions). History features come from all of a participant's reported conflicts. The ridge regression uses 31 features (number of conflicts, strategies per episode, repertoire size, mean and SD of success, mean conflict strength, and the 25 label rates), with 5-fold cross-validation over participants repeated 20 times (mean and SD over repeats). Partial ρ controls the ranks of mean success, strategies per episode, and repertoire size.

Predictor	Knowledge	Regulation
Mean reported success	.234	−.007
Strategies per episode	.190	−.002
Repertoire size	.168	.090
Ridge regression, 31 features	.245 (SD .024)	.158 (SD .030)
LLMs, median (maximum)	.178 (.235)	.003 (.096)
LLMs, median partial ρ (IQR)	−.023 (−.042 to −.003)	–

Cross-person overlap. For each model–language condition, we average the Jaccard overlap between the predicted sets of every pair of persistence episodes from two different participants, skipping pairs in which both sets are empty; this average weights episode pairs equally. At P4 the median over the 44 conditions is .547 (range .408–.839; P1: .587), against .100 for the same pairs of reports; every condition exceeds the report value. A model's set for one person therefore overlaps more with its set for another person than with the person's own report (median $J = .142$), and adding the P2–P4 blocks lowers the median cross-person overlap from .587 at P1 to .547 at P4. Because Jaccard overlap rises with set size, two checks address size. Permuting each label across episodes, which keeps every label's rate, gives a median of .540 for the predictions at P4 (observed .547) and .112 for the reports (observed .100), so the cross-person overlap of the predictions follows from label rates concentrated on a common set of labels, and the reports overlap less than their label rates imply. In a count of distinct sets, which larger sets make more numerous, the 1,135 predicted sets take a median of 312.5 distinct forms at P4 (range 104–639; 211 at P1), against 602 for the reports, fewer than the reports in 42 of 44 conditions (44 of 44 at P1). Predictions for two episodes of the same participant, whose prompts share the persona block, overlap more than predictions for different participants (median .743 vs. .547 at P4, all 44 conditions), as reports do (.225 vs. .100).

Task 3 success tracking. Between-person and within-person correlations separate two ways of reducing error: ranking people by their typical success and tracking which of a person's episodes went better (Table 14). From P1 to P4, the between-person correlation rises in persistence and initiation, whereas the within-person correlation changes little and stays below the ridge regression trained on the episode context and reported strategies. Both ridge regressions standardize the features within each training fold and select the penalty from 25 log-spaced values (10^{-2}–10^{4} for Task 3, 10^{-1}–10^{4} for Task 4) by efficient leave-one-out cross-validation on the training fold. The Task 3 model uses 63 features (25 strategy indicators, conflict strength, and 37 activity and goal indicators, one of which marks a missing goal domain) and five participant-grouped folds (seed 240725); the Task 4 model uses 20 repeats of 5-fold cross-validation (seed 240725), sets the SD of success to 0 for the 32 participants with fewer than two success ratings, and fills the remaining missing features (one participant's mean success and, for ten participants with only inhibition reports, the rates of three labels absent from the inhibition checklist) with the median over participants.

Task 4 references. A single history feature, the person's mean reported success, ranks Knowledge above 43 of the 44 LLM conditions, and the models' Knowledge predictions correlate with this feature (median $\rho = .446$) and with strategies per episode (.464), and their Regulation predictions correlate with the same two features at .343 and .516. Controlling for these two features and repertoire size removes the models' Knowledge association (Table 15). The ridge regression ranks Regulation above every LLM condition. The Knowledge–Regulation difference in the LLM correlations has a 95% participant-bootstrap interval above zero in 36 of 44 conditions (1,000 resamples per condition).

Other checks. On the 661 episodes with earlier reports, the LLM median Jaccard is .158 (range .114–.197), against .257 for the earlier-report predictor. Task 2 scoring includes the five persistence episodes without a success rating; excluding them changes the median Task 2 Jaccard from .1430 to .1428.

B Evaluation Protocol Details

This appendix gives the protocol details for Section 3.

B.1 Dataset Fields

Source structure. The evaluation uses the German experience-sampling study analyzed by Bürgler & Hennecke (2024) (503 participants, eight prompts a day for ten days), whose data Wenzel et al. (2024) also analyzed. Participants completed baseline questionnaires and then reported self-control conflicts in daily life. The study presented two questionnaire forms at random; the evaluation uses the conflicts reported with the form that recorded strategies. The analysis links each conflict episode to the participant-level baseline measures used for persona conditioning. The data include 3,725 conflicts from 468 participants, a median of seven conflicts per participant (interquartile range

4–11); Table 16 gives the counts by conflict type. The success-scored subset contains 3,707 conflicts and is used to score Task 3.

Table 16: Self-control conflict types in the evaluation set. The descriptions follow the study codebook and summarize the situation participants reported.

Type	Situation represented	Episodes
Initiation	Starting an aversive or effortful activity in order to reach a goal or avoid negative consequences.	1,909 total; 1,898 success-scored
Persistence	Continuing an aversive or effortful activity in order to reach a goal or avoid negative consequences.	1,135 total; 1,130 success-scored
Inhibition	Resisting a temptation despite wanting to give in to it.	681 total; 679 success-scored

The field inventory was defined by aligning the evaluation variables with the original study codebook and the persona-item inventory. Tables 17, 18, 19, and 20 document, respectively, the source fields with derived scores, the strategy indicators, the episode-level question stems, and the persona blocks used for conditioning, scoring, subgroup analyses, or diagnostics. Table 21 documents how psychometric persona items are rendered in prompts.

Persona-item rendering. For P2–P4, psychometric fields are rendered at the item level. Each prompt unit pairs the original questionnaire stem with the participant's verbal Likert response label. This design gives the model semantically interpretable persona evidence while keeping numeric source scores separate for scoring, aggregation, and evaluation.

Strategy scoring references. The strategy reference is a participant-reported binary set defined over checklist items. Conflict-specific source strategy columns are mapped into 25 merged indicators. Applicability masks evaluate model predictions against the strategies on the relevant conflict type's checklist, the options the participant chose from.

Scoring references for Tasks 3–4. The self-regulatory success score is a 1–7 participant-reported score, with higher values indicating more successful self-regulation. It is recorded for 3,707 episodes, which are used to score Task 3; Tasks 1 and 2 are scored on all 3,725 strategy-labeled episodes. Task 4 compares participant-history predictions against participant-level MISCS Knowledge and Regulation composites. The Knowledge composite is the arithmetic mean of six items covering understanding one's strengths and weaknesses, remembering past conflicts, using situation-specific strategies, knowing which strategies one uses, compensating for weaknesses with strengths, and judging how well one resolves conflicts. The Regulation composite is the arithmetic mean of six items covering strategy consideration and choice, analyzing strategy usefulness during conflicts, monitoring progress during conflicts, evaluating goal accomplishment afterward, and checking whether all strategies were considered. The codebook marks the response scale as 1–5 agreement and does not indicate reverse-keying for these 12 selected items. Across 468 participants, both source composites are finite for 467 participants. Knowledge has mean 3.345, SD 0.675, and range 1.333–5.000; Regulation has mean 2.879, SD 0.760, and range 1.000–5.000.

B.2 Task Schemas With Prompt Figures

Task rationale. The evaluation defines four prediction tasks from the self-report data. In each task, one component of the participant report is used as input and another component is evaluated. Task 1 predicts the participant-reported strategy set from persona and episode context. Tasks 2 and 3 form opposite directions of the strategy–success relation: Task 2 uses observed success to predict the strategy set, whereas Task 3 uses the reported strategy set to predict subjective success. Task 4 aggregates repeated conflict reports into a participant history and predicts MISCS Knowledge and Regulation without providing the MISCS item responses as input. The four tasks evaluate whether models track reported strategy choice, the relation between appraisal and strategy, and participant-level metacognitive self-control.

Table 17: Source information, derived scores, and scoring references. The table lists fields used for task conditioning, scoring, subgroup analyses, or diagnostics; the complete source codebook contains additional measures outside this evaluation. Source question wording is taken from the study codebook.

Field family	Source question or item family	Role in this evaluation
Demographics	Age; sex; marital status; education; current status such as employed or student.	P1 persona block; participant-level subgroup diagnostics.
Conflict type	Initiation, persistence, or inhibition conflict label.	Defines task scope and strategy applicability masks.
Conflict strength	"How strong did you experience this inner conflict?" on a 1–7 scale.	Episode-level process measure and exact-level diagnostic subgroup.
Persistence activity	Persistence activity categories such as self-study, preparing a meal, housework, shopping, sports, or social contact.	Main persistence context for episode-level prediction.
Persistence goal	Goal-domain fields, including academic/work, health/fitness, financial, daily life, leisure, social, and other goals.	Main persistence context for why the participant continued the activity.
Strategy checklist	Multiple-selection strategy items, e.g., adding something positive, changing the activity, thinking about positive consequences, planning, monitoring, removing distractions, suppression, commitment.	Task 1 and Task 2 scoring reference; Task 3 conditioning variable; Task 4 participant-history input.
Subjective self-regulatory success	Merged participant-reported success score; higher values indicate more successful self-regulation.	Task 2 conditioning variable; Task 3 scoring reference; Task 4 participant-history input.
Big Five	The 15 BFI-2-XS items covering Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness (Soto & John, 2017).	P2 persona block.
Trait self-control	Capacity for self-control items covering inhibition, initiation, and continuation (Hoyle & Davisson, 2016); the broader trait self-control tradition motivates the domain relevance (Tangney et al., 2004; de Boer et al., 2011).	P3 persona block and self-control subgroup bands.
MISCS	Metacognition in Self-Control items adapted from metacognitive-awareness constructs (Schraw & Dennison, 1994; Bürgler et al., 2022).	P4 persona block for Tasks 1–3; excluded from Task 4 conditioning information and used as participant-level comparison scores.

Table 18: Self-regulatory strategy labels used for set prediction. Conflict-type masks restrict scoring to strategies available for the relevant source checklist.

Strategy label	Codebook description
Task enrichment	Add something positive to make the activity more pleasant.
Change activity	Change the activity itself.
Change feelings	Change how one feels.
Change environment	Change the environment in which the activity is performed or the temptation is encountered.
Positive consequences	Think about positive consequences of acting in line with the goal.
Distraction	Shift attention away from the activity or temptation.
Substance use	Use a substance such as coffee or an energy drink.
Finish goal	Remind oneself that the episode will soon be over.
Planning	Make a plan or set a specific time.
Goal setting	Define a goal or subgoals.
Reappraisal	Think differently about the activity or temptation.
Monitoring	Check goal progress.
Negative consequences	Think about negative consequences of not acting in line with the goal.
Process focus	Focus on the activity itself and how it is being performed.
Remove distractions or temptation	Reduce or remove distractions or temptations.
Self-reward	Promise or give oneself a later reward.
Self-talk	Talk to oneself for motivation.
Social support	Draw on support from other people.
Suppression	Force oneself to begin, continue, or resist.
Self-discipline training	Train self-discipline outside the focal activity or temptation (variable `outside_help`).
Public pressure	Put oneself under time, social, or performance pressure.
Argue with self	Argue with oneself about why the goal-consistent action should be taken.
Commitment	Remind oneself of a commitment made to oneself or others.
Affirmation	Remind oneself that one can do it.
Other strategy	Report another strategy.

Table 19: Episode-level question stems used for task conditioning, scoring, or auditing. The table prints source-question stems for readability. Evaluation procedures use the participant's observed answer or the corresponding strategy list.

Construct	Codebook stem or item family	Evaluation use
Conflict strength	"How strong did you experience this inner conflict?"	Episode-level 1–7 process measure and exact-level subgroup axis.
Persistence activity	"Which category or categories are the most likely to be assigned to the activity you wanted to persist in?" with source categories such as commuting, work, self-study, preparing a meal, shopping, housework, sports, and social contact.	Concrete persistence activity field; this field motivates the main-results focus.
Persistence importance	"How important was it to you to persist in the activity?"	Episode-level importance measure when observed for the participant episode.
Goal presence	"Did you pursue one or more goals with the activity you wanted to persist in?"	Goal-context field indicating whether continuation served an explicit goal.
Goal domain	"Which area or areas can these goals be assigned to?" with academic/work, health/fitness, financial, daily-life, leisure, social, spiritual/religious, and other domains.	Goal-domain field used to represent why continuation mattered.
Strategy use	"Did you use one or more of the following strategies to persist in this activity?" (Hennecke et al., 2019; Wenzel et al., 2024).	Task 1 and Task 2 multi-label scoring reference; Task 3 and Task 4 inputs.
Strategy items	Source checklist items include enriching the activity, changing the activity, changing feelings, changing the environment, thinking about positive consequences, planning, monitoring, removing distractions, self-reward, self-talk, social support, suppression, public commitment, and reminding oneself of a commitment.	Mapped into the 25-indicator strategy ontology with conflict-type applicability masks.
Success estimate	"How well did you manage to persist in this activity?"	Component of the 1–7 success score.
Success satisfaction	"How satisfied are you with how well you persisted in the activity?"	Component of the merged self-regulatory success score.
Aversiveness descriptors	Items asking whether the activity was unpleasant, physically effortful, mentally effortful, monotonous, pointless, emotion-eliciting, boring, or frustrating.	Source context retained for auditing the protocol's field selection.

Table 20: Persona information blocks used in P2–P4. This table records construct, item count, and source; Table 21 shows prompt-style narrative examples of the item-level rendering.

Block	Item count	Construct represented	Source
P2 Big Five	15	Stable personality differences used to test whether broad trait information changes episode-level prediction.	Soto & John (2017)
P3 trait self-control	20	Inhibition, initiation, and continuation capacities aligned with the three conflict types.	de Boer et al. (2011); Hoyle & Davisson (2016); Tangney et al. (2004)
P4 MISCS Knowledge	6	Knowledge about one's own self-control strengths, weaknesses, strategies, and situation dependence.	Schraw & Dennison (1994); Bürgler et al. (2022)
P4 MISCS Regulation	6	Planning, monitoring, strategy evaluation, and post-conflict self-evaluation.	Schraw & Dennison (1994); Bürgler et al. (2022)

Table 21: Narrative examples of persona-item rendering in P2–P4. Each prompt unit is written as an interpretable item–response phrase. The examples show prompt-style strings from the item inventory; the final response label is filled with each participant's verbal Likert label.

Persona block	Rendered narrative item–response examples
P2 Big Five	*Tends to be quiet*: mostly applies. *Is compassionate, has a soft heart*: applies completely. *Tends to be disorganized*: mostly does not apply. *Worries a lot*: neutral. *Is original, comes up with new ideas*: mostly applies.
P3 Inhibition capacity	*I am able to resist temptations*: applies completely. *I have trouble resisting my cravings*: rather does not apply. *I can deny myself something I want but do not need*: applies. *When I want something that is bad for me, I go after it anyway*: does not apply at all.
P3 Initiation capacity	*I waste a lot of time before getting down to work*: rather does not apply. *I delay as long as possible before starting something unpleasant*: neutral. *Even when the list of things to do is long, it is easy for me to get started*: applies. *I get started on new projects right away*: applies completely.
P3 Continuation capacity	*I am able to keep doing what I think I should do even when other people would stop*: applies. *When I commit to doing something difficult, I see it through to the end*: applies completely. *I find it hard to continue doing something I do not want to do*: rather does not apply.
P4 MISCS Knowledge	*I understand my strengths and weaknesses when dealing with self-control conflicts*: agree. *I use different regulatory strategies depending on the situation*: strongly agree. *I am aware of what strategies I use to resolve self-control conflicts*: agree.
P4 MISCS Regulation	*I consider several strategies to resolve a self-control conflict before I try to resolve it*: strongly agree. *I find myself pausing regularly to check how well I am resolving a self-control conflict*: neutral. *I ask myself how well I accomplished my goals once I have resolved a self-control conflict*: agree.

Figures 3, 4, 5, and 6 print the English user-prompt templates of Tasks 1–4 used by the prompt generator. Braced fields denote participant-, episode-, or list-specific values substituted before inference; German prompts use parallel translations of the same templates. No separate system prompt was used.

Prompt for Task 1: Strategy Set Prediction

You are a person with the following characteristics:
- Age: {AGE}
- Sex: {SEX}
- Marital status: {MARITAL_STATUS}
- Education: {EDUCATION}
- Currently: {CURRENT_STATUS}

Your personality (Big Five Inventory):
{TRAIT}:
"{BFI_ITEM_TEXT}": {VERBAL_LIKERT_RESPONSE}

Your self-control capacity:
{INHIBITION/INITIATION/CONTINUATION}:
"{HOYLE_ITEM_TEXT}": {VERBAL_LIKERT_RESPONSE}

Your metacognitive self-control profile:
Metacognitive knowledge (understanding your own self-control patterns):
"{MISCS_KNOWLEDGE_ITEM}": {VERBAL_AGREEMENT_RESPONSE}
Metacognitive regulation (planning, monitoring, and evaluating your self-control):
"{MISCS_REGULATION_ITEM}": {VERBAL_AGREEMENT_RESPONSE}

You need to continue {ACTIVITY}, but you don't feel like doing it.
This activity serves your: {GOAL_DOMAIN}.
The conflict feels {STRENGTH_WORD} to you ({STRENGTH} on a 1-7 scale).

Below is a numbered list of self-regulatory strategies. Select the numbers of ALL strategies you would actually use in this situation.

1. {STRATEGY_NAME} — {STRATEGY_DESCRIPTION}
{K}. {STRATEGY_NAME} — {STRATEGY_DESCRIPTION}

Respond with ONLY the selected numbers as a JSON array (e.g., `[1, 9, 13]`).

Figure 3: English prompt template for Task 1: strategy-set prediction. The rendered prompt contains the item-response lines for the active persona regime and the conflict-type-specific strategy list.

Prompt for Task 2: Success-Conditioned Strategy Prediction

{PERSONA_CARD}

You need to continue {ACTIVITY}, but you don't feel like doing it.
This activity serves your: {GOAL_DOMAIN}.
The conflict feels {STRENGTH_WORD} to you ({STRENGTH} on a 1-7 scale).

In this situation, you managed to handle the conflict {SUCCESS_WORD} ({SUCCESS} on a 1-7 scale, where 1 = not well at all and 7 = very well).

Given how well you handled this conflict, which self-regulatory strategies did you actually use? Select the numbers of ALL strategies you used.

1. {STRATEGY_NAME} — {STRATEGY_DESCRIPTION}
{K}. {STRATEGY_NAME} — {STRATEGY_DESCRIPTION}

Respond with ONLY the selected numbers as a JSON array (e.g., `[1, 9, 13]`).

Figure 4: English prompt template for Task 2: success-conditioned strategy prediction. The {PERSONA_CARD} block is the same generated persona card shown in Figure 3.

Figures 3–5 show persistence prompts, and Figure 6 shows the Task 4 history template, which lists conflicts of all three types; the activity and goal-domain fields are specific to persistence episodes (Section 3.1). Initiation and inhibition prompts follow the same persona and strategy-list structure;

Prompt for Task 3: Strategy-Conditioned Success Prediction

{PERSONA_CARD}

You experienced a self-control conflict: you had to continue {ACTIVITY}, but you didn't feel like doing it. This activity served your: {GOAL_DOMAIN}. The conflict felt {STRENGTH_WORD} to you ({STRENGTH} on a 1-7 scale).

To handle this conflict, you used the following strategies:
- {STRATEGY_NAME} — {STRATEGY_DESCRIPTION}
- {STRATEGY_NAME} — {STRATEGY_DESCRIPTION}

How successful did you feel in handling this conflict?
Rate on a scale from 1 (not at all successful) to 7 (very successful).
Respond with ONLY a single integer from 1 to 7.

Figure 5: English prompt template for Task 3: prediction of the participant's subjective success from the reported strategy set.

Prompt for Task 4: Participant-History MISCS Prediction

You are a person with the following demographics:
- Age: {AGE}
- Sex: {SEX}
- Marital status: {MARITAL_STATUS}
- Education: {EDUCATION}
- Currently: {CURRENT_STATUS}

Over recent days, you experienced the following self-control conflicts:
  Episode {J} ({CONFLICT_TYPE}): conflict strength={STRENGTH}/7, success={SUCCESS}/7
  Strategies used:
    - {STRATEGY_NAME} — {STRATEGY_DESCRIPTION}
    - {STRATEGY_NAME} — {STRATEGY_DESCRIPTION}

Based on your experience with these self-control conflicts, rate how much you agree with each of the following statements about yourself on a scale from 1 (strongly disagree) to 5 (strongly agree).

1. I understand my strengths and weaknesses when dealing with self-control conflicts.
2. I am good at remembering past self-control conflicts.
3. I use different regulatory strategies depending on the situation.
4. I am aware of what strategies I use to resolve self-control conflicts.
5. I use my strengths to compensate for my weaknesses when dealing with daily self-control conflicts.
6. I am a good judge of how well I resolve self-control conflicts.
7. I consider several strategies to resolve a self-control conflict before I try to resolve it.
8. I think of several strategies to resolve a self-control conflict and choose the best one.
9. I find myself analyzing the usefulness of strategies during self-control conflicts.
10. I find myself pausing regularly to check how well I am resolving a self-control conflict.
11. I ask myself how well I accomplished my goals once I have resolved a self-control conflict.
12. I ask myself if I have considered all strategies after I resolve a self-control conflict.

Respond with ONLY a JSON array of 12 integers (e.g., `[4, 3, 5, 4, 3, 4, 2, 3, 2, 2, 3, 2]`).

Figure 6: English prompt template for Task 4: participant-history MISCS prediction. The rendered prompt repeats the episode block for every available participant episode and withholds MISCS item responses.

their conflict block names the conflict type (starting or resisting) and conflict strength, and the strategy list follows the conflict-type applicability mask.

B.3 Method Details

Model panel. The analysis uses completions from 22 models (Table 22). Task 1 contains 29,800 prompts per model: 3,725 strategy-labeled episodes × four persona regimes × two languages. The Task 1 persistence subset contains 9,080 prompts per model: 1,135 persistence episodes × four regimes × two languages. Tasks 2 and 3 were also sent for all 3,725 episodes (29,800 prompts per model each). Task 3 is scored on the 3,707 success-scored episodes (9,040 persistence prompts from 1,130 episodes), and Task 2 on all episodes, including five persistence episodes without a success rating (Appendix A.7). For the 18 episodes without a success rating (11 initiation, five persistence, two inhibition), the Task 2 prompt states the scale midpoint, "moderately (4 on a 1-7 scale, ...)", and the Task 4 history prints the success of such an episode as "not reported". Task 4 contains 936 prompts per model: 468 participants × two languages.

Table 22: Model panel, in the order used in every figure and table: closed models, then open-weight models; families in alphabetical order; within a family, larger to smaller (total parameters for open-weight models, provider tier for closed models, and the GPT-5.4 series before GPT-5). Identifiers are the API or checkpoint names used for the runs.

Access	Family	Model	Identifier
Closed	Claude	Claude Sonnet 4.6	`claude-sonnet-4-6`
Closed	Claude	Claude Haiku 4.5	`claude-haiku-4-5-20251001`
Closed	Gemini	Gemini 3 Flash	`gemini-3-flash-preview`
Closed	Gemini	Gemini 3.1 Flash Lite	`gemini-3.1-flash-lite-preview`
Closed	GPT	GPT-5.4	`gpt-5.4`
Closed	GPT	GPT-5.4 mini	`gpt-5.4-mini`
Closed	GPT	GPT-5.4 nano	`gpt-5.4-nano`
Closed	GPT	GPT-5	`gpt-5`
Closed	GPT	GPT-5 mini	`gpt-5-mini`
Closed	GPT	GPT-5 nano	`gpt-5-nano`
Open-weight	DeepSeek	DeepSeek V3.1	`DeepSeek-V3.1`
Open-weight	Gemma	Gemma 4 31B	`gemma-4-31B-it`
Open-weight	Gemma	Gemma 4 26B-A4B	`gemma-4-26b-a4b-it`
Open-weight	GLM	GLM 4.7	`GLM-4.7`
Open-weight	Llama	Llama 4 Scout	`Llama-4-Scout-109B-Instruct`
Open-weight	Llama	Llama 3.3 70B	`Llama-3.3-70B-Instruct`
Open-weight	Llama	Llama 3.1 8B	`Llama-3.1-8B-Instruct`
Open-weight	Llama	Llama 3.2 3B	`Llama-3.2-3B-Instruct`
Open-weight	Qwen	Qwen 3.5 122B-A10B	`Qwen3.5-122B-A10B-FP8`
Open-weight	Qwen	Qwen 3.5 35B-A3B	`Qwen3.5-35B-A3B`
Open-weight	Qwen	Qwen 3.5 27B	`Qwen3.5-27B`
Open-weight	Qwen	Qwen 3.5 9B	`Qwen3.5-9B`

Persona information regimes. P1 is a minimal participant description containing demographics and life-status fields. P2 adds Big Five item responses because broad personality dimensions are a standard representation of stable individual differences (John et al., 1991). P3 adds trait self-control responses because initiation, continuation, and inhibition directly index the domain capacities underlying the three conflict types (de Boer et al., 2011; Hoyle & Davisson, 2016; Tangney et al., 2004). P4 adds MISCS responses because metacognitive knowledge and regulation should be directly relevant to monitoring, selecting, and evaluating self-control strategies (Schraw & Dennison, 1994; Bürgler et al., 2022). The regimes are cumulative so that P1, P2, P3, and P4 can be evaluated as ordered information conditions; adjacent contrasts localize the contribution of each added block, and the full P4−P1 contrast summarizes the overall movement from minimal to richest participant information.

Prompt languages. English and German prompts use identical participant identifiers, episode identifiers, persona regimes, and scoring scripts. The source study materials provide German wording and English translations for the questionnaire items used in the prompts; prompt construction follows these provided wordings. Language analyses therefore test sensitivity to the prompt language under fixed reports.

Decoding settings. Inference used deterministic or near-greedy decoding wherever the model interface exposed the relevant controls. The purpose of these settings was to keep repeated prompt conditions comparable and to reduce stochastic variation in model outputs, so that observed differences across persona regimes, tasks, and languages reflect the evaluated prompt information. Table 23 records the sampling controls used for each provider family. Each API received the controls it accepts; for APIs that treat a zero nucleus value inconsistently, a very small nonzero top-p value was used to enforce near-greedy sampling. Requests to the open-weight models carried an empty system message.

Table 23: Decoding controls used to reduce output stochasticity. Dashes denote controls left at the endpoint default, as the endpoint requires or by design.

Model group	Temperature	Top-p
Claude Sonnet 4.6, Claude Haiku 4.5	0	–
Gemini 3 Flash, Gemini 3.1 Flash Lite	0	0
GPT-5.4, GPT-5.4 mini, GPT-5.4 nano	0	10^{-9}
GPT-5, GPT-5 mini, GPT-5 nano	–	–
Open-weight models (served endpoints)	0	10^{-9}

Output parsing. API errors were retried (up to five attempts per request, with failed batch requests re-submitted), and each prompt's returned response was scored once. For Tasks 1–2, the parser removes Markdown code fences and reads the JSON array, else the first bracketed list, else English strategy names in the text; each number selects the option at that position in the conflict-type-specific list shown in the prompt, numbers outside the list and unmatched names are discarded, and a repeated option counts once. A response without a valid option is scored as an empty predicted set (160 Task 1 and 2,347 Task 2 responses of 655,600 each; 2,358 of these 2,507 are explicit empty lists). For 114 Llama 3.2 3B German responses without a usable list, the integers from 1 to the list length that appear in the text are read as the selected options. For Task 3, the first number in the response is clipped to $[1, 7]$, and a response without a number is scored as 4 (six Llama 3.2 3B refusals and one empty Claude Haiku 4.5 response); two further Llama 3.2 3B refusals, whose first number belongs to a telephone number, are clipped to 7. For Task 4, the parser reads the JSON array, else the first bracketed list when it holds exactly 12 numbers; the one response without such a list (a Llama 3.2 3B German refusal) receives no score, and two Llama 3.1 8B German vectors that contain a 0 enter the subscale means as returned.

B.4 Metric Definitions

Let G_e be the participant-reported strategy set for episode e, and let $\hat{G}_{e,m,r,l}$ be the strategy set predicted by model m under persona regime r and prompt language l. Jaccard overlap is

$$J(e, m, r, l) = \frac{|G_e \cap \hat{G}_{e,m,r,l}|}{|G_e \cup \hat{G}_{e,m,r,l}|}.$$

Precision is $|G_e \cap \hat{G}_{e,m,r,l}|/|\hat{G}_{e,m,r,l}|$ and is undefined when the predicted set is empty. Recall is $|G_e \cap \hat{G}_{e,m,r,l}|/|G_e|$ and is undefined when the reported set is empty. Per-episode F1 is the harmonic mean of precision and recall. When exactly one of two compared sets is empty, Jaccard and F1 are 0 and the undefined precision or recall is omitted from the participant mean; when both sets are empty, a prediction scores Jaccard 1 and F1 0 against its report, and a pair of two reports or two predictions is omitted from the pair averages. The reference comparisons (Tables 2 and 13) score the precision of an empty predicted set as 0; with 23 empty predicted sets across the 44 P4 persistence conditions,

both rules give a median precision of .178 (at most .003 apart in any condition). Predicted set size is $|\hat{G}_{e,m,r,l}|$, signed set-size error is $|\hat{G}_{e,m,r,l}| - |G_e|$, and absolute set-size error is $||\hat{G}_{e,m,r,l}| - |G_e||$.

Label-specific direction. For individual strategy label k, the model prediction rate under condition (m, r, l) is

$$\text{rate}_k(m, r, l) = \frac{1}{|E_k|} \sum_{e \in E_k} \mathbf{1}[k \in \hat{G}_{e,m,r,l}],$$

where E_k is the set of episodes for which label k falls inside the conflict-type applicability mask. The participant-report rate is computed analogously from G_e. Label direction is the sign of the prediction rate minus the report rate: positive values indicate over-selection and negative values indicate under-selection relative to the participant sample. For each label k within condition (m, r, l), significance is assessed with an exact binomial McNemar test (two-sided) on the discordant episode pairs: episodes where model and participant disagree on label k are classified as false positives (model selects, participant does not) or false negatives (participant selects, model does not), and the null hypothesis is that false-positive and false-negative counts are equally likely. Holm correction is applied within the per-task, per-label family (all model–language–conflict-type–regime conditions for a given label within a task, up to 528) at $\alpha = 0.05$. A label carries corrected support for over- or under-selection when the Holm-adjusted p-value is below 0.05 and the sign of the rate difference indicates the stated direction. Label-direction counts in Section 4.1 are Holm-significant conditions, except the matched-cardinality count (32 of 40 conditions), which counts raw signs. The label-direction diagnostic complements aggregate Jaccard by resolving which specific regulatory actions drive the overlap gap: two conditions with the same Jaccard can differ in whether the gap comes from missing reported labels or from inserting unreported labels, and label direction separates these sources.

Participant-level check. Episodes are nested in participants, and the McNemar test treats episode pairs as independent. Each label test is therefore repeated at the participant level: for each condition, a two-sided Wilcoxon signed-rank test compares the participants' mean differences between predicted and reported selection of the label with zero (zero differences dropped), with Holm correction within the same families. Of the 5,628 label–condition pairs with Holm-significant McNemar tests at P4 (Tasks 1–2, all conflict types), 5,100 (90.6%) remain Holm-significant with the same sign, and none reverses. In Task 1 persistence, 854 of 951 remain (89.8%), including self-talk (42 of 44 conditions), positive consequences (43), commitment (42), removing temptations (40), goal setting (39), planning (39), changing feelings (33), changing the activity (32), and distraction (23). Self-discipline training keeps its under-selection in initiation (37 of 44 conditions; 32 reports from 20 participants) and loses it in persistence (24 reports from 15 participants) and inhibition (16 reports from 13 participants; 0 of 44 each).

Matched-cardinality slice. The matched-cardinality slice restricts comparison to episodes where $|\hat{G}_{e,m,r,l}| = |G_e|$, holding predicted set size equal to the participant-reported set size. Models that list more strategies than participants report produce higher Jaccard denominators and mechanically lower overlap; the matched-cardinality restriction removes this confound. If over- and under-selection of specific labels persist when the predicted and reported sets have identical cardinality, the label bias holds at fixed output length. The matched-cardinality episode sample varies across models and conditions; median slice sizes are reported alongside the direction results.

Labels with a prespecified direction. Ten strategy labels were given a prespecified direction before the analysis. Five were expected to be over-selected on the basis that internal-cognitive self-regulation labels are over-represented in psychology and self-help corpora: self-talk, positive consequences, reappraisal, commitment, and goal setting. Five were expected to be under-selected on the basis that they require coordination with the physical or social environment: self-discipline training (variable `outside_help`, defined in the codebook as practice outside the focal activity; its expected direction was set from the variable name, which suggests help from others), substance use, social support, change environment, and public pressure. These ten labels are used in the direction summary of Section 4.1. Figure 1(c) shows the six over-selected and four under-selected labels with the most Holm-significant conditions, selected from the results to display the most recurrent directions; Figures 7 and 8 report all 25 labels. In Task 1 persistence, all five labels expected to be

over-selected are over-selected with Holm significance in most conditions. Of the five expected to be under-selected, only self-discipline training is, in episode-level tests (the participant-level check confirms none of these conditions): social support and changing the environment are over-selected in more conditions than under-selected, and substance use and public pressure are split (Table 34).

For Task 3, let y_e be participant-reported success and $\hat{y}_{e,m,r,l}$ be model-predicted success. Absolute error is $|\hat{y}_{e,m,r,l} - y_e|$ and signed error is $\hat{y}_{e,m,r,l} - y_e$. The reported success y_e is the mean of the two 1–7 success items and ends in .5 for 1,499 of the 3,707 scored episodes (466 of the 1,130 in persistence); every scored model prediction is an integer from 1 to 7, as the prompt requests, so the absolute error on these episodes is at least 0.5. Episode-level metrics for Tasks 1–3 are averaged within participant before inference.

Participant-pair Jaccard reference. The participant-pair Jaccard reference quantifies the strategy-overlap scale that arises between two unrelated participants drawn from the same conflict-type sample, independently of any model output. For each pair of episodes (e_i, e_j) reported by distinct participants under the same conflict type, both with a nonempty reported set ($|G_{e_i}| > 0$ and $|G_{e_j}| > 0$),

$$J_{\text{pair}}(e_i, e_j) = \frac{|G_{e_i} \cap G_{e_j}|}{|G_{e_i} \cup G_{e_j}|}.$$

The distribution of J_{pair} over all such pairs provides the population-level reference, computed over the 496,453 different-participant pairs among the 999 persistence reports that list at least one strategy. If a model's persona-conditioned Jaccard sits inside the participant-pair distribution, the model operates at the overlap scale of two unrelated participants. The reference therefore separates two interpretive frames: relative improvement (higher Jaccard under richer persona information) and absolute person–episode recovery (Jaccard above the participant-pair upper quartile, Q75). Summary statistics (mean, median, interquartile range) and the pair count are reported for persistence (Table 38, Figure 10).

For Task 4, the model returns an ordered 12-item MISCS vector $(z_1, \ldots, z_{12})$ on the 1–5 response scale. The first six positions correspond to MISCS Knowledge and the final six positions correspond to MISCS Regulation, matching the fixed MISCS item order used by the prompt generator. Predicted subscale scores are arithmetic means,

$$\widehat{K}_{m,l,p} = \frac{1}{6}\sum_{i=1}^{6} z_i, \qquad \widehat{R}_{m,l,p} = \frac{1}{6}\sum_{i=7}^{12} z_i,$$

for model m, language l, and participant p. The source comparison scores are the participant-level MISCS Knowledge and Regulation composites, both on the same 1–5 scale. A Task 4 item-level prediction is considered complete when the audit recovers all 12 values (Appendix B.3 notes two Llama 3.1 8B German vectors that contain a 0). Main Task 4 score analyses use finite predicted and source subscale scores. The subscale scores are computed from the first bracketed list of 12 numbers in each response, and the audit recovers an item trace only from a response that consists of this list or contains no other number; a condition in which some responses surround the list with explanatory text containing numbers is marked as an aggregate-score condition, and in each such response (246 of 466 for Llama 3.2 3B German, 19 for Claude Sonnet 4.6 English, 22 for Llama 4 Scout English, 9 for Llama 4 Scout German) the subscale scores are the means of the first six and last six values of that list, all within 1–5. Spearman ρ is computed between predicted and source MISCS scores within each model–language–subscale combination. Task 4 MAE and signed bias are reported as calibration diagnostics. For model m and language l, Task 4 Knowledge MAE is $\text{MAE}_K(m,l) = \frac{1}{P}\sum_{p=1}^{P} |\widehat{K}_{m,l,p} - K_p|$ and signed bias is $\text{Bias}_K(m,l) = \frac{1}{P}\sum_{p=1}^{P}(\widehat{K}_{m,l,p} - K_p)$, where K_p is the source composite and P is the number of participants with finite predicted and source scores; Regulation MAE and bias are defined analogously from $\widehat{R}_{m,l,p}$ and R_p.

Subscale reliability. Task 4 compares model predictions against MISCS-12 Knowledge and Regulation composites. Both are six-item subscales whose reliability sets how much measurement error in the source composite attenuates observed rank associations; the indices below quantify this attenuation, which is reported alongside the rank results.

Internal consistency is assessed with three complementary indices. When item errors are uncorrelated, Cronbach's α equals reliability under essential tau-equivalence (equal true-score loadings) and

is a lower bound to reliability when loadings differ. McDonald's ω is computed from a one-factor maximum-likelihood factor analysis of the standardized items as $\omega = (\sum_i \lambda_i)^2/[(\sum_i \lambda_i)^2 + \sum_i \psi_i]$, with loadings λ_i and uniquenesses ψ_i; it allows unequal loadings and estimates reliability under the congeneric one-factor model with uncorrelated errors. Rank-based α, Cronbach's α computed on within-item ranks, checks sensitivity to item-level non-normality; this study treats agreement with Cronbach's α within 0.02 as robustness to the marginal item distributions.

Unidimensionality is summarized by the share of total standardized item variance carried by the first eigenvalue of each subscale's item correlation matrix, with 40% as a diagnostic threshold chosen for this study.

Under classical test theory, unreliability in the source composite attenuates the observed Spearman ρ between predicted and source subscale scores. The outcome-only correction $\rho_{\text{corrected}} = \rho_{\text{observed}}/\sqrt{\alpha}$ estimates the association between the predicted score and the error-free source score under two assumptions: the predicted score is error-free, and source measurement error is uncorrelated with it. The correction is derived for Pearson correlations and is applied here to Spearman's ρ. Corrected values are reported as sensitivity diagnostics alongside the observed associations. The correction uses Cronbach's α; because α does not exceed ω of the standardized items for either subscale (Table 27), dividing by $\sqrt{\alpha}$ gives the larger of the two corrected values.

B.5 Analysis Protocol

Persona-regime inference. Tasks 1–3 evaluate the full P1–P4 sequence of information regimes. The main persistence subset includes 353 participants with strategy-labeled persistence episodes for Tasks 1–2 and 351 participants with success-scored persistence episodes for Task 3; these counts reflect the participants in the 468-participant sample who contributed at least one persistence episode to the corresponding evaluation set. The inferential contrasts are P2−P1, P3−P2, P4−P3, and P4−P1 within participant, model, conflict type, prompt language, task, and metric. Paired Wilcoxon signed-rank tests are used because the paired metric differences are bounded, frequently tied, and often non-normal. The analysis inspects paired-difference histograms, Q-Q diagnostics, Shapiro–Wilk diagnostics within the test's valid sample-size range, tie counts, zero-difference counts, and finite-pair counts before reporting each family. Each test is two-sided and ranks the nonzero within-participant paired differences.

Task 4 uses Spearman rank tests with null hypothesis H_0: $\rho = 0$ (no monotonic association between predicted and source subscale scores), tested via the t-approximation with $n-2$ degrees of freedom, where n is the participant count per condition (466–467). MAE and signed bias are reported after rank association as calibration diagnostics. History-length analyses regress Task 4 absolute score error and signed bias separately on log episode count with prespecified covariates (sex, employment, student status, and the three self-control capacity bands, all drawn from the P1–P3 persona blocks and independent of the withheld MISCS target) and use OLS with HC3 (heteroscedasticity-consistent, variant 3) covariance estimates because participant-level heteroscedasticity and influential cases are possible. Residual-vs-fitted plots, Q-Q plots, scale-location diagnostics, hat-value summaries, and Cook's-distance summaries audit these regressions.

Multiplicity correction. Holm correction is applied within each prespecified metric or subscale family at $\alpha = 0.05$ (Holm, 1979). BH q-values are retained as diagnostic columns (Benjamini & Hochberg, 1995). Adjacent persona contrasts localize where movement occurs across the ordered information regimes, while P4−P1 summarizes the full change from minimal to richest participant information.

Complete-item-trace sensitivity subset. The complete-item-trace subset for Task 4 uses the model–language conditions with complete item-level parsed MISCS vectors, in which the audit recovers all 12 predicted item values, so that each predicted subscale score traces to its parsed items. The subset confirms that the Knowledge–Regulation contrast is preserved under the stricter provenance requirement.

Paired effect sizes. Paired effect sizes for the persona-regime contrasts are reported as rank-biserial correlations r_{rb} (Kerby, 2014) from the Wilcoxon signed-rank statistic, defined as

$$r_{rb} = \frac{2(T^+ - T^-)}{n(n+1)},$$

where T^+ and T^- are the positive and negative rank sums over the n nonzero paired differences. Values near zero indicate negligible standardized movement; conditions with $|r_{rb}| > 0.20$, a descriptive threshold chosen for this study, are counted in Table 65. Bracketed r_{rb} values are interquartile ranges across conditions. Both r_{rb} and condition-count summaries are reported so that recurrence across conditions and magnitude within a condition are both visible.

TOST equivalence tests. For the Task 4 rank associations of both subscales, equivalence is assessed with two one-sided tests (TOST; Lakens, 2017). Each Spearman ρ ($n = 466$–467 participants per condition) is transformed to $z = \text{arctanh}(\rho)$ with standard error $1/\sqrt{n-3}$, and equivalence at the margin $\Delta = 0.10$ on the Fisher-z scale ($|\rho| < \tanh 0.10 = .0997$), a margin chosen for this study, is declared when both one-sided tests reject $\zeta \leq -\Delta$ and $\zeta \geq +\Delta$ at $\alpha = 0.05$, where $\zeta = \text{arctanh}(\rho_{\text{pop}})$. Sensitivity margins of ± 0.05 and ± 0.15 are reported alongside the ± 0.10 margin.

JZS Bayes factors. For each model–language condition, BF_{01} is the JZS Bayes factor (Rouder et al., 2009) with Cauchy prior scale $r = \sqrt{2}/2$, computed from the t statistic of Spearman's ρ, $t = \rho\sqrt{(n-2)/(1-\rho^2)}$, and sample size n; the panel summary is the median BF_{01} over the 44 conditions. Values above 3 indicate at least moderate evidence for the null (above 10, strong); values below $1/3$ indicate at least moderate evidence for the alternative; values between $1/3$ and 3 are anecdotal. The Bayes-factor diagnostic is reported alongside TOST because the two address different questions: TOST assesses whether the data support population associations within the prespecified bounds, and the Bayes factor quantifies the relative evidence for $\rho = 0$ and for the alternative without a prespecified margin.

Cross-language consistency. Cross-language directional agreement between English and German model-level P4−P1 movements is assessed with Pearson correlation over the 22 models. The correlation quantifies whether the same models that move favorably under persona blocks in one language also move favorably in the other. Direction-agreement counts—the number of models whose P4−P1 sign matches across both languages—are reported alongside the correlation to separate concordance in sign from concordance in magnitude.

B.6 Subgroup Analyses

Subgroup analyses evaluate heterogeneity across prespecified participant and episode strata: eight participant-level axes and the episode-level conflict-strength axis (Table 24). Axes were admitted only when they used groups coded in the source data, instrument midpoints, or exact observed process values; the axes and the level-size minima were prespecified in the analysis plan. The minimum level size is 30 for standard axes, 50 for sensitive demographic axes, and 10 nonzero paired differences for within-subgroup paired tests. Levels below these minima are skipped. Omnibus tests use Wilcoxon rank-sum (Mann–Whitney U) for two-group axes and Kruskal–Wallis for multi-group axes (conflict strength, which has exact observed levels 1–7), with Holm correction applied within each axis-by-task family of 132 tests.

Table 24: Prespecified subgroup axes and rationale. Axes use groups coded in the source data, instrument midpoints, or exact observed process levels.

Axis	Split rule	Rationale
Sex	Female vs. male; other and missing excluded from inferential comparisons.	Demographic stratum coded in the source data, with sensitive-axis minimum $n = 50$.
Current employment	Employed vs. not employed.	Current life context may affect everyday conflict exposure and persona interpretation; status flag coded in the source data.
University-student status	Student vs. non-student.	The sample contains many students, and persistence activities include academic goal contexts.
Self-control continuation	> 4 vs. ≤ 4 on the 1–7 instrument midpoint.	Continuation capacity is directly aligned with persistence conflicts (Hoyle & Davisson, 2016; de Boer et al., 2011).
Self-control inhibition	> 4 vs. ≤ 4 on the 1–7 instrument midpoint.	Inhibition capacity is relevant to temptation-resistance conflicts (Tangney et al., 2004; Hoyle & Davisson, 2016).
Self-control initiation	> 4 vs. ≤ 4 on the 1–7 instrument midpoint.	Initiation capacity is relevant to starting aversive activities (de Boer et al., 2011; Hoyle & Davisson, 2016).
MISCS Knowledge	> 3 vs. ≤ 3 on the 1–5 midpoint.	Persona-sensitivity stratum for Tasks 1–3; kept separate from Task 4 subgroup inference, whose comparison score it forms.
MISCS Regulation	> 3 vs. ≤ 3 on the 1–5 midpoint.	Metacognitive-regulation stratum for Tasks 1–3; kept separate from Task 4 subgroup inference, whose comparison score it forms.
Conflict strength	Exact observed levels 1–7; missing excluded.	Process-strength diagnostic.

C Units of Analysis With Correction Families

C.1 Units of Analysis

The analyses separate four units: participant-level paired tests inside a model-language condition, summaries over the fixed 22-model-by-2-language panel, participant-report distance diagnostics, and Task 4 participant-rank analyses. The separation keeps descriptive panel recurrence, corrected participant-level tests, and participant-report recovery in their correct denominators.

For Tasks 1–2, positive signed movement means higher Jaccard overlap with participant-reported strategies. For Task 3, positive signed movement means lower absolute success error. Participant-report distance is computed against the observed participant report: $1 - \text{Jaccard}$ for Tasks 1–2 and absolute success error for Task 3. Task 4 rank entries use Spearman association between predicted and reported MISCS subscale scores. Across participants, Cronbach α is .719 for Knowledge and .786 for Regulation (Table 27). The Regulation analysis in Section 4.4 tests whether the models' ratings from histories rank participants on the subscale, which a ridge regression on the same histories does.

When paired Wilcoxon effects are reported, r_{rb} denotes the matched-pairs rank-biserial correlation. The sign is oriented so that favorable movement is positive: higher Jaccard for Tasks 1–2 and lower absolute success error for Task 3.

Table 25 gives the primary numerical summary as the compact index for the detailed appendix rows. The supporting sections preserve denominator, correction-family, and prompt-block detail behind the finding that the models fail to reproduce the process.

C.2 Primary Summary With Denominators

The prompt-block definition fixes the interpretation of P1–P4. P1 contains demographic and life-status fields. P2 adds Big Five responses. P3 adds trait self-control responses. P4 adds MISCS Knowledge and Regulation item responses. Adjacent contrasts are therefore content contrasts across psychological blocks. Task 4 withholds the target MISCS item responses and uses conflict histories as the predictor. The sample definition also anchors persistence as the main Tasks 1–3 analysis because persistence episodes record the activity being continued and its goal (Section 3.1).

The task denominators are metric-specific (Table 26). Tasks 1–2 use 353 persistence participants with analyzable strategy reports. Task 3 uses the 351 of these participants whose records include the success-appraisal target that the paired absolute success-error test scores. Task 4 uses the 466–467 participants per model-language condition with a finite source composite (467) and a finite predicted score in that condition (466–467) on the tested subscale (Knowledge or Regulation); the stricter complete-six-item reliability analysis uses 464 (Knowledge) and 463 (Regulation) participants (Table 27).

C.3 Correction Families

Table 28 defines the count units, and Table 29 records the multiplicity family attached to each corrected decision. Direction counts over model-language conditions describe recurrence across the model panel; corrected tests are participant-level tests within the stated family, except for label-direction tests, which use episode-level McNemar tests within the per-task, per-label family. The all-conflict-type Tasks 1–3 families provide multiplicity control, while the main narrative reports the persistence subset.

Table 25: Primary persistence evidence summary. Counts summarize model-language conditions unless the row names a descriptive reference such as the participant-pair overlap; bracketed intervals are interquartile ranges across model-language conditions.

Result	Quantity	Panel evidence	Magnitude
P4 strategy-set breadth	Task 1 and Task 2 labels/output	above participant mean in 88/88 conditions	participant mean 3.099; P4 ranges 5.608–17.039 / 4.262–13.490
Task 1 strategy-label direction	P4 prediction rate − participant rate	Self-talk 44/44; Self-discipline training 41/44; Holm: Self-talk 44/44; Self-discipline training 41/44 (participant-level: 42/44 and 0/44)	Self-talk +0.832; Self-discipline training −0.021
Task 2 strategy-label direction with success	P4 prediction rate − participant rate	Self-talk 43/44; Substance use 38/44; Holm: Self-talk 42/44; Substance use 35/44	Self-talk +0.594; Substance use −0.107
P4 matched-cardinality sensitivity	selected label rates after matched set size	median direction retained in the 10 displayed task–label rows	Task 1 exact rows 34.5 [20.0, 53.2]; Task 2 59.5 [38.8, 89.2]
Descriptive Jaccard baselines	persistence reports (pairs: nonzero reports)	participant pair 0.128; size-matched diagnostics random 0.092 and in-sample frequency 0.272 (both use the reported set size; Appendix A.1)	LLM P4 medians: Task 1 0.142; Task 2 0.143
Task 1	Jaccard overlap	P4 direction +25/−19; Holm +3/−3	Δ +0.001 [−0.002, +0.006]; distance 0.858
Task 2	Jaccard overlap	P4 direction +27/−17; Holm +7/−6	Δ +0.003 [−0.005, +0.010]; distance 0.857
Task 3: strategy to success	absolute success error	P4 direction +35/−9; Holm +9/−0	Δ +0.055 [+0.007, +0.093]; distance 1.306
Task 3 P4−P3 block contrast	absolute success error	P4 direction +13/−31; Holm +1/−2	Δ −0.014 [−0.043, +0.002]

Table 26: Participant and episode denominators. Task columns report the participant count used in paired participant-level tests; episode counts are raw human records by conflict type.

Conflict type	Participants	Episodes	Task 1 N	Task 2 N	Task 3 N
Persistence	353	1,135	353	353	351
Initiation	426	1,909	426	426	426
Inhibition	276	681	276	276	276
Task 4 rank analytic conditions	466–467	3,725	–	–	–

Table 27: MISCS-12 internal consistency across participants on the six scored items per subscale (Knowledge: MISCS 4, 11, 12, 18, 20, 21; Regulation: MISCS 2, 15, 19, 23, 25, 26). McDonald's ω is computed from a one-factor maximum-likelihood factor analysis of the standardized items; rank-based α is Cronbach α on within-item ranks, an ordinal robustness check. The first-eigenvalue share reports the proportion of total item variance carried by the first principal component as a unidimensionality diagnostic.

Subscale	Participants	Cronbach α	McDonald's ω	Rank-based α	First-eigenvalue share
Knowledge	464	0.719	0.721	0.711	41.8%
Regulation	463	0.786	0.786	0.780	48.4%

Table 28: Count units used in the Results and supporting analyses.

Count	Unit	Definition	Location
22	fixed model panel	the 22 models of Table 22	all analyses
468	human participant pool	participants with at least one conflict episode	sample description
3,725	human episode pool	all conflict episodes in the analysis file	sample description
353/1,135	persistence participants/episodes	participant count / episode count	main Tasks 1–3
426/1,909	initiation participants/episodes	participant count / episode count	conflict-type analyses
276/681	inhibition participants/episodes	participant count / episode count	conflict-type analyses
44	single-conflict-type model-language conditions	22 models $\times$ 1 conflict type $\times$ 2 languages	Tasks 1–3
132	all-conflict-type model-language conditions	22 models $\times$ 3 conflict types $\times$ 2 languages	cross-conflict-type summaries
44	Task 4 subscale conditions	22 models $\times$ 2 languages per MISCS subscale	Task 4 rank tests
44	Task 4 adjusted conditions	22 models $\times$ 2 languages per subscale-outcome	Task 4 history-length diagnostics
40/44	matched-cardinality conditions	Task 1 / Task 2 retained model-language conditions	matched-cardinality slice
264	language comparisons	22 models $\times$ 3 conflict types $\times$ 4 prompt blocks	language scope audit
3,564	subgroup omnibus families	9 prespecified axes $\times$ 396 families	supporting audit

Table 29: Correction-family record for fixed analyses. Columns give the multiplicity scope, contrast, and correction denominator used by each family. Family names are written as analysis labels.

Analysis family	Scope	Contrast / metric	k and correction
Task 1 action overlap	Task 1; all three conflict types	P4−P1; Jaccard overlap	132; Holm over model-language-conflict-type conditions; BH audit
Task 2 action overlap	Task 2; all three conflict types	P4−P1; Jaccard overlap	132; Holm over model-language-conflict-type conditions; BH audit
Task 3 absolute success error	Task 3; all three conflict types	P4−P1; absolute success error	132; Holm over model-language-conflict-type conditions; BH audit
Task 3 signed success error	Task 3; all three conflict types	P4−P1; signed success error	132; Holm over model-language-conflict-type conditions; BH audit
Persona-regime contrasts	Tasks 1–3; all conflict types	Adjacent and full prompt-block contrasts	12 task/metric/contrast families; each $k = 132$; Holm within family
Strategy-label direction	Tasks 1–2; all conflict types, labels, and prompt blocks	Model prediction rate vs. participant-report rate	one family per task and label over all model–language–conflict-type–regime conditions; $k = 352$ or 528; Holm within family
Task 4 Knowledge rank	Task 4; Knowledge subscale	History-predicted score vs. participant score; Spearman ρ	44; Holm within subscale; BH audit
Task 4 Regulation rank	Task 4; Regulation subscale	History-predicted score vs. participant score; Spearman ρ	44; Holm within subscale; BH audit
Task 4 history-length error	Task 4; outcome by subscale	Log history length; absolute or signed score error	44 per outcome-by-subscale family; Holm within family; BH audit
Subgroup P4−P1 diagnostics	Tasks 1–3; prespecified subgroup axes	Within-subgroup P4−P1 metric deltas	8,691 tests; Holm within task/axis/metric; BH audit
Subgroup omnibus tests	Tasks 1–3; prespecified subgroup axes	Omnibus subgroup families	3,564 families; 72 Holm-significant and meeting the level-size minima
Subgroup post-hoc comparisons	Tasks 1–3; families meeting the level-size minima	Final post-hoc and all-pairs audit	66 final tests; 3,168 all-pairs comparisons
Task 4 subgroup diagnostics	Task 4; prespecified participant-level subgroup axes	Knowledge and Regulation subgroup diagnostics	1,056 tests; 0 Holm-significant

D EXAMPLE EPISODE WITH STRATEGY GROUPS

The main example (Table 30) is included because aggregate overlap hides which participant labels are missed and which labels are substituted. The example gives the concrete object of evaluation: recovery of the reported strategies under a fixed ontology.

Table 30: Example of an episode mismatch (GPT-5.4 nano, German, P4, Task 1).

Component	Strategy labels
Evaluation condition	GPT-5.4 nano, German, P4, Task 1, episode 2051
Participant report	Change environment; Substance use
Correct overlap	none
Missed participant labels	Change environment; Substance use
Predicted labels absent from the report	Self-talk; Commitment; Planning; Goal setting; Positive consequences; Reappraisal; Monitoring; Process focus; Affirmation; Argue with self; Self-reward; Remove distractions or temptation; Social support; Public pressure; Suppression; Finish goal

Table 31: Strategy groups used for label-direction summaries.

Group	n	Strategy labels
Internal	9	Self-talk; Commitment; Positive consequences; Negative consequences; Reappraisal; Process focus; Affirmation; Argue with self; Self-reward
Planning	3	Planning; Goal setting; Monitoring
Environmental	2	Remove distractions or temptation; Change environment
Social	2	Social support; Public pressure
Self-discipline	1	Self-discipline training (variable `outside_help`)
In-the-moment adjustments	5	Change activity; Change feelings; Distraction; Substance use; Suppression
Goal-focused	2	Finish goal; Task enrichment
Other	1	Other strategy

The ontology in Table 31 defines the strategy groups used in the prose. Internal self-regulation labels and planning labels are common regulatory actions; in-the-moment, environmental, and social labels are the labels most directly tied to the immediate episode; self-discipline training refers to practice outside the focal activity. The grouping is applied to the fixed model outputs.

The example fixes the evaluation target. A prediction can be semantically fluent and still fail the observed recovery criterion when the participant-reported labels are absent. For that reason, the main Results measure person–episode fidelity by recovery of the participant report.

E LABEL DIRECTION UNDER SET-SIZE CONTROLS

The first Results claim rests on three linked diagnostics: P4 strategy-set breadth, label-specific substitution, and a matched-cardinality sensitivity slice. The appendix retains the full numerical counterpart to Figure 1.

E.1 SET SIZE

The set-size evidence shows that P4 predicts broader repertoires than participants report in every persistence model-language condition (Table 32). Broad sets make raw Jaccard harder to interpret, so the main Results combine set size with label direction before interpreting the recovery gap.

E.2 LABEL DIRECTION

The main evidence is the alignment between rate direction and strategy content (Table 33). Internal labels such as Self-talk and Commitment are repeatedly over-selected, while labels that adjust feel-

Table 32: Persistence P4 strategy-set size (labels per episode; participant-balanced; 44 model–language conditions).

Task	Participant mean	P4 median	P4 range	Above participant mean	English	German
Task 1	3.099	8.689	5.608–17.039	44/44	22/22	22/22
Task 2	3.099	6.465	4.262–13.490	44/44	22/22	22/22

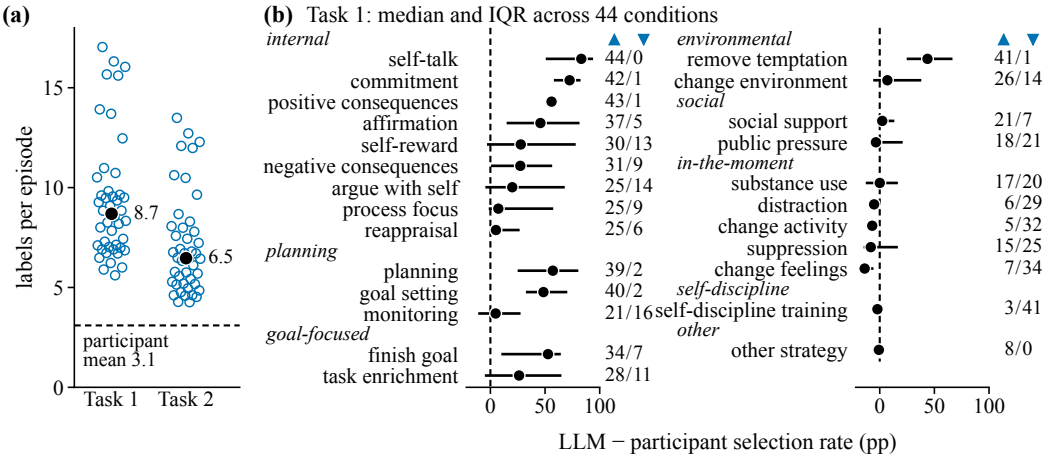

Figure 7: Across the 44 model–language conditions, models over-select internal and planning labels and under-select adjustments of feelings and activity (Task 1, persistence, P4). (a) Predicted strategies per episode in Tasks 1 and 2 for each condition (hollow circles) with medians (filled) and the participants' mean (dashed). (b) Selection rate minus participants' report rate for all 25 labels: median (circle) and interquartile range (bar) over conditions; the numbers give Holm-significant over- and under-selected conditions (columns marked ▲ and ▼). Holm counts are episode-level McNemar tests; the participant-level check does not confirm the self-discipline training direction in persistence or inhibition (Appendix B.4).

Table 33: Persistence P4 strategy-label tests. Direction is the model prediction rate relative to the participant report rate. Holm counts are episode-level McNemar tests; the participant-level check does not confirm the self-discipline training direction in persistence or inhibition (Appendix B.4).

Task	Strategy label	Direction	Direction conditions	Holm conditions	Median rate Δ [IQR]
Task 1	Self-talk	over	44/44	44/44	+0.832 [+0.507, +0.940]
Task 1	Positive consequences	over	43/44	43/44	+0.559 [+0.530, +0.569]
Task 1	Commitment	over	42/44	42/44	+0.725 [+0.581, +0.825]
Task 1	Remove distractions or temptation	over	42/44	41/44	+0.437 [+0.247, +0.666]
Task 1	Self-discipline training	under	41/44	41/44	−0.021 [−0.021, −0.021]
Task 2	Self-talk	over	43/44	42/44	+0.594 [+0.242, +0.809]
Task 2	Positive consequences	over	42/44	42/44	+0.503 [+0.435, +0.533]
Task 2	Self-discipline training	under	41/44	41/44	−0.021 [−0.021, −0.021]
Task 2	Commitment	over	41/44	40/44	+0.648 [+0.374, +0.771]
Task 2	Substance use	under	38/44	35/44	−0.107 [−0.128, −0.061]

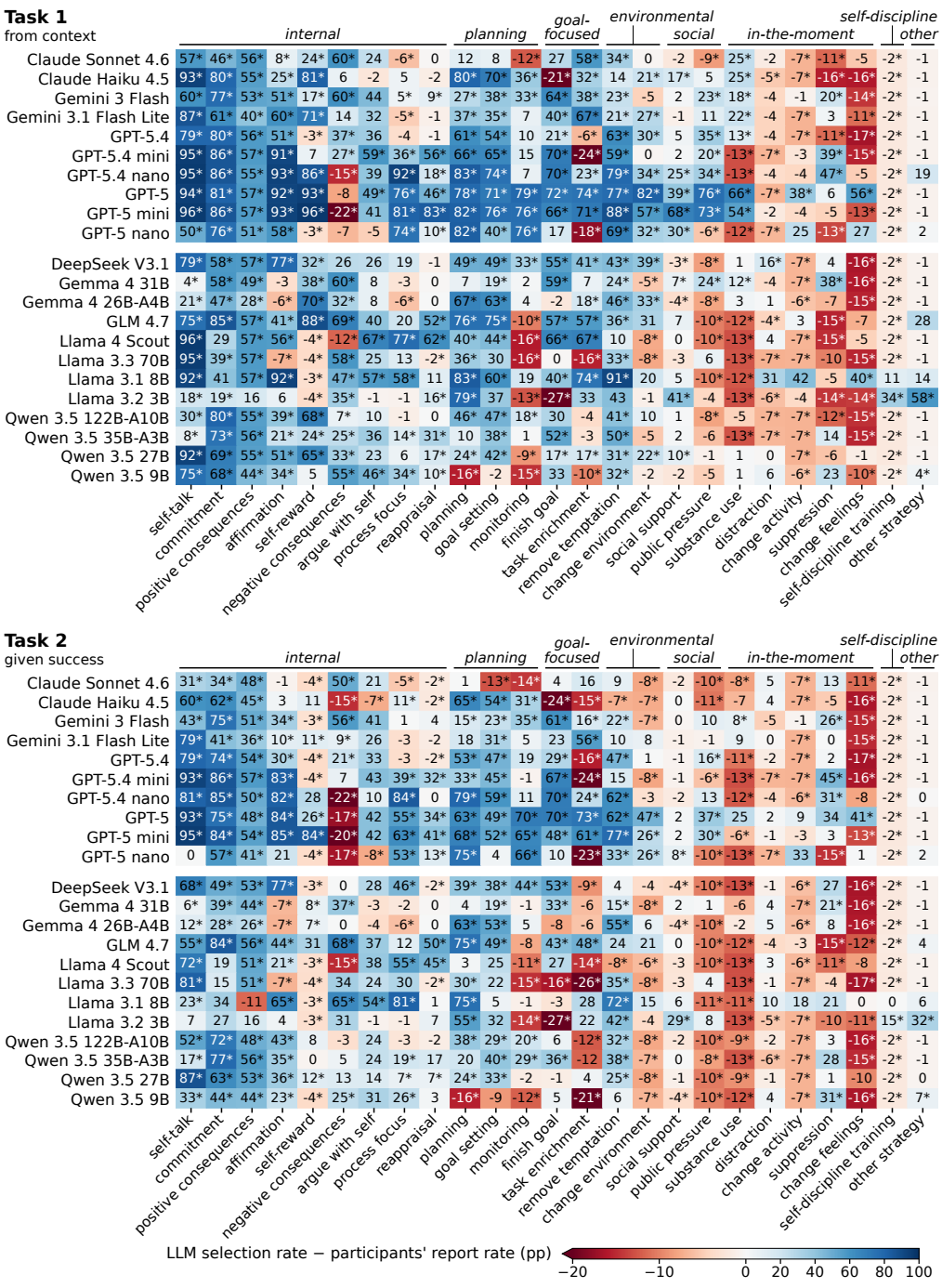

Figure 8: Most models (18–22 of 22) over-select self-talk, commitment, and positive consequences in both tasks (persistence, P4), and 20 under-select self-discipline training in episode-level tests, which the participant-level check does not confirm (Appendix B.4). Cells give the LLM selection rate minus the participants' report rate (percentage points), averaged over each model's English and German conditions, for all 25 labels in Task 1 (strategies from context, top) and Task 2 (given the reported success, bottom); labels are grouped as in Table 31 and ordered by the Task 1 median, as in Figure 7. $*$: both language conditions Holm-significant (McNemar) in the same direction. Negative values carry a minus sign; colors are clipped at -20. Rows follow the model order of Table 22. Self-discipline training denotes training self-discipline outside the activity; Figure 1(c) shows ten of these labels.

ings or the activity in the moment (Change feelings, Change activity, Distraction) and Self-discipline training are under-selected, the last in episode-level tests only (Appendix B.4); Substance use is under-selected in Task 2 but split in Task 1. The full rows (Table 34) preserve the numerical basis for the label directions described in Section 4.1. Among labels that act on the situation, Remove distractions or temptation is over-selected in 41/44 conditions. The substitution is therefore systematic across the ontology, and labels that act on the situation appear among both the over-selected and the under-selected labels.

Figure 7 summarizes the full direction pattern, including less prominent labels, and Figure 8 gives the values by model.

E.3 Matched-Cardinality Slice

The matched-cardinality slice (Table 35) is a post-generation sensitivity analysis. We first generated the fixed outputs under the standard prompts, then selected the episode rows where the model's predicted set size equaled the participant's reported set size; this selection compares predicted and reported label rates at equal set size. The retained subset preserves selected label directions at matched set size, so it is consistent with the main label-direction interpretation from the episode-level McNemar tests and their participant-level check (Appendix B.4).

F Persona-Block Contrasts

Prompt-block movement, participant-report distance, and signed success-error movement quantify different parts of the same persistence result. Together, the detailed rows show how Task 3 improvement coexists with a large participant-report distance.

F.1 Block Contrasts

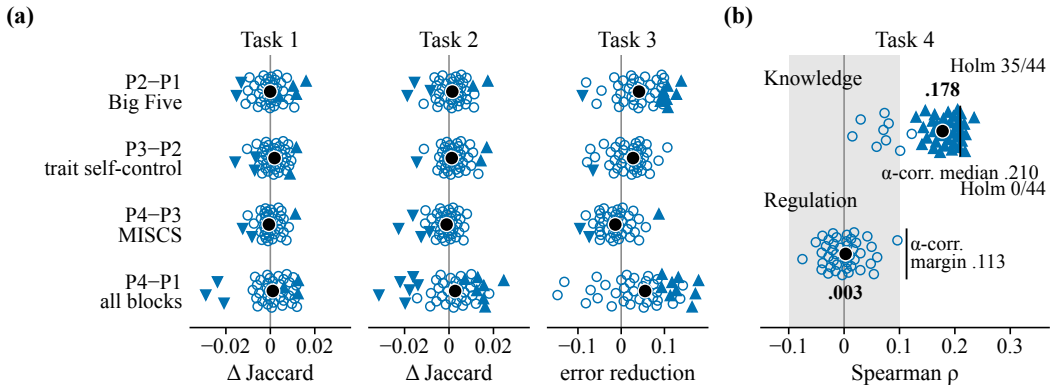

Figure 9: Single persona blocks change predicted success, and strategy-recovery changes stay near zero (persistence); outcome-reliability correction leaves the Task 4 contrast in place. (a) Change in Task 1–2 Jaccard and reduction in Task 3 error when each block is added (P2−P1 Big Five, P3−P2 trait self-control, P4−P3 MISCS) and for all three blocks (P4−P1); hollow circles are conditions without Holm significance, triangles Holm-significant favorable (▲) or unfavorable (▼) conditions, filled circles medians. (b) Task 4 Spearman ρ for all conditions with the $\pm.10$ band; the Knowledge tick marks the median divided by $\sqrt{\alpha}$ ($.178/\sqrt{.719} = .210$), and the Regulation tick marks the equivalence margin divided by $\sqrt{\alpha}$ ($.10/\sqrt{.786} = .113$), the margin on the outcome-reliability-corrected scale.

The block contrasts show content-specific movement across P1–P4 (Table 36, Figure 9). The P4−P1 Task 3 contrast is often favorable, and P4 lowers predicted success relative to P1 in 19 of 22 models (Figure 11). The P4−P3 contrast can move the other way after MISCS responses are added to the trait self-control block. The contrast therefore depends on the content added and the baseline used.

Table 34: Full persistence P4 strategy-label audit across the fixed 25-label ontology. Rates are proportions. Signed Δ is model prediction rate minus participant-report rate; the Holm cell reports episode-level McNemar over- and under-selection counts over 44 model-language conditions; the participant-level check does not confirm the self-discipline training direction (Appendix B.4).

Group	Strategy label	Participant rate	T1 P4	T1 Δ; Holm +/−	T2 P4	T2 Δ; Holm +/−
internal	self-talk	0.043	0.875	+0.832; +44/−0	0.637	+0.594; +42/−1
internal	commitment	0.141	0.866	+0.725; +42/−1	0.789	+0.648; +40/−2
planning	planning	0.166	0.737	+0.571; +39/−2	0.579	+0.414; +34/−2
planning	goal setting	0.192	0.678	+0.485; +40/−2	0.552	+0.360; +35/−7
internal	positive consequences	0.430	0.989	+0.559; +43/−1	0.933	+0.503; +42/−2
internal	negative consequences	0.219	0.493	+0.274; +31/−9	0.278	+0.059; +22/−13
internal	reappraisal	0.021	0.072	+0.051; +25/−6	0.038	+0.017; +19/−12
planning	monitoring	0.161	0.209	+0.048; +21/−16	0.219	+0.058; +22/−13
internal	process focus	0.061	0.133	+0.073; +25/−9	0.169	+0.108; +26/−11
internal	affirmation	0.070	0.528	+0.458; +37/−5	0.369	+0.299; +33/−8
internal	argue with self	0.087	0.287	+0.200; +25/−14	0.098	+0.011; +19/−14
internal	self-reward	0.038	0.316	+0.278; +30/−13	0.010	−0.028; +16/−24
environmental	remove distractions or temptation	0.083	0.520	+0.437; +41/−1	0.347	+0.264; +34/−5
environmental	change environment	0.078	0.148	+0.070; +26/−14	0.014	−0.065; +10/−27
social	social support	0.041	0.063	+0.023; +21/−7	0.033	−0.008; +9/−14
self-discipline	self-discipline training	0.021	0.000	−0.021; +3/−41	0.000	−0.021; +3/−41
social	public pressure	0.107	0.073	−0.035; +18/−21	0.016	−0.092; +11/−28
in-the-moment	change activity	0.070	0.001	−0.069; +5/−32	0.000	−0.069; +3/−38
in-the-moment	change feelings	0.166	0.028	−0.138; +7/−34	0.008	−0.157; +4/−37
in-the-moment	distraction	0.070	0.019	−0.052; +6/−29	0.059	−0.011; +7/−14
in-the-moment	substance use	0.130	0.130	+0.000; +17/−20	0.022	−0.107; +4/−35
in-the-moment	suppression	0.157	0.075	−0.082; +15/−25	0.207	+0.050; +21/−15
goal-focused	finish goal	0.275	0.802	+0.527; +34/−7	0.515	+0.240; +29/−14
goal-focused	task enrichment	0.259	0.522	+0.263; +28/−11	0.171	−0.088; +16/−23
other	other strategy	0.008	0.000	−0.008; +8/−0	0.000	−0.008; +7/−0

Table 35: Persistence P4 strategy-label direction after matched-cardinality filtering.

Task	Strategy label	Over conditions	Under conditions	Tie conditions	Median rate Δ [IQR]	Exact rows per condition
Task 1	Self-talk	32/40	7/40	1/40	+0.403 [+0.092, +0.758]	34.5 [20.0, 53.2]
Task 1	Positive consequences	30/40	6/40	4/40	+0.124 [+0.019, +0.246]	34.5 [20.0, 53.2]
Task 1	Commitment	33/40	6/40	1/40	+0.246 [+0.121, +0.468]	34.5 [20.0, 53.2]
Task 1	Remove distractions or temptation	27/40	11/40	2/40	+0.199 [−0.022, +0.367]	34.5 [20.0, 53.2]
Task 1	Self-discipline training	3/40	31/40	6/40	−0.036 [−0.059, −0.017]	34.5 [20.0, 53.2]
Task 2	Self-talk	35/44	7/44	2/44	+0.411 [+0.089, +0.642]	59.5 [38.8, 89.2]
Task 2	Positive consequences	34/44	7/44	3/44	+0.188 [+0.051, +0.273]	59.5 [38.8, 89.2]
Task 2	Commitment	36/44	6/44	2/44	+0.323 [+0.178, +0.500]	59.5 [38.8, 89.2]
Task 2	Self-discipline training	3/44	34/44	7/44	−0.030 [−0.046, −0.012]	59.5 [38.8, 89.2]
Task 2	Substance use	5/44	39/44	0/44	−0.161 [−0.198, −0.122]	59.5 [38.8, 89.2]

F.2 Distance From Reports

Participant-report distance gives the recovery counterpart to relative prompt movement. The comparison is P4 against the participant report. The median distances are 0.858 (Task 1) and 0.857 (Task 2) on the Jaccard scale and 1.306 on the 1–7 success scale (Table 37), against median P4−P1 movements of $+0.001$, $+0.003$, and $+0.055$ (Table 36), so a favorable relative movement leaves the predictions far from the participant's own reported strategies or success value.

The participant-pair reference puts the small Jaccard movements in scale (Table 38, Figure 10). Within nonzero persistence-episode pairs from the participant report, the participant-pair Jaccard distribution has mean 0.128, median 0.083, and interquartile range $[0.000, 0.200]$. LLM panel P4 medians of 0.142 (Task 1) and 0.143 (Task 2) sit within ± 0.02 of this mean (over pairs in which at least one report lists a strategy, the pair mean is .100, below all 22 model means; Figure 10); the highest model mean is Gemma 4 31B (Task 1: 0.172 vs. Q75 = 0.200; Task 2: 0.174 vs. Q75 = 0.200), followed by Claude Sonnet 4.6 and Gemini 3 Flash, and no model reaches the participant-pair upper quartile in either task. Because raw Jaccard is set-size-sensitive, the reference serves as a scale diagnostic.

F.3 Signed Success Error

Signed error locates the direction of the Task 3 change (Table 39). P4 often lowers predicted success relative to P1 (Figure 11). The shift can reduce absolute success error when P1 predicts higher success than participants report.

Table 36: Persistence prompt-block contrasts with magnitude. Positive signed values mean higher Jaccard for Tasks 1–2 and lower absolute success error for Task 3.

Task	Metric	Contrast	Conditions	Holm	Med. signed Δ [IQR]	Med. signed r_{rb} [IQR]
Task 1	Jaccard overlap	P2−P1	+21/−23	+4/−2	−0.000 [−0.004, +0.003]	−0.012 [−0.103, +0.073]
Task 1	Jaccard overlap	P3−P2	+26/−18	+2/−3	+0.002 [−0.002, +0.006]	+0.004 [−0.067, +0.107]
Task 1	Jaccard overlap	P4−P3	+19/−25	+1/−2	−0.001 [−0.003, +0.004]	+0.000 [−0.051, +0.132]
Task 1	Jaccard overlap	P4−P1	+25/−19	+3/−3	+0.001 [−0.002, +0.006]	+0.040 [−0.103, +0.134]
Task 2	Jaccard overlap	P2−P1	+27/−17	+2/−3	+0.001 [−0.003, +0.005]	+0.046 [−0.059, +0.138]
Task 2	Jaccard overlap	P3−P2	+28/−16	+2/−1	+0.001 [−0.001, +0.005]	+0.055 [−0.068, +0.118]
Task 2	Jaccard overlap	P4−P3	+18/−26	+0/−4	−0.001 [−0.004, +0.002]	−0.006 [−0.106, +0.069]
Task 2	Jaccard overlap	P4−P1	+27/−17	+7/−6	+0.003 [−0.005, +0.010]	+0.043 [−0.092, +0.162]
Task 3: strategy to success	absolute success error	P2−P1	+35/−9	+8/−1	+0.041 [+0.015, +0.093]	+0.143 [+0.034, +0.286]
Task 3: strategy to success	absolute success error	P3−P2	+31/−13	+0/−1	+0.027 [−0.006, +0.040]	+0.079 [−0.032, +0.186]
Task 3: strategy to success	absolute success error	P4−P3	+13/−31	+1/−2	−0.014 [−0.043, +0.002]	−0.053 [−0.187, +0.011]
Task 3: strategy to success	absolute success error	P4−P1	+35/−9	+9/−0	+0.055 [+0.007, +0.093]	+0.101 [+0.013, +0.245]

Table 37: Persistence P4 participant-report distance: 1 − Jaccard for Tasks 1–2 and absolute success error (1–7 scale) for Task 3; medians and interquartile ranges over the 44 model-language conditions.

Conflict	Task	Primary metric	Median distance	Distance IQR
Persistence	Task 1	Jaccard overlap	0.858	[0.845, 0.864]
Persistence	Task 2	Jaccard overlap	0.857	[0.846, 0.872]
Persistence	Task 3: strategy to success	Absolute success error	1.306	[1.243, 1.348]

Table 38: Descriptive persistence Jaccard baselines; the participant-pair row uses nonzero participant-reported strategy sets, and the LLM rows use all 1,135 episodes.

Comparison	Units	Definition	Mean	Median
Participant-pair overlap	all pairs	pairs of nonzero persistence reports	0.128	0.083
LLM P4 Task 1	44 conditions	median over model-language conditions of participant-balanced Jaccard	–	0.142
LLM P4 Task 2	44 conditions	median over model-language conditions of participant-balanced Jaccard	–	0.143

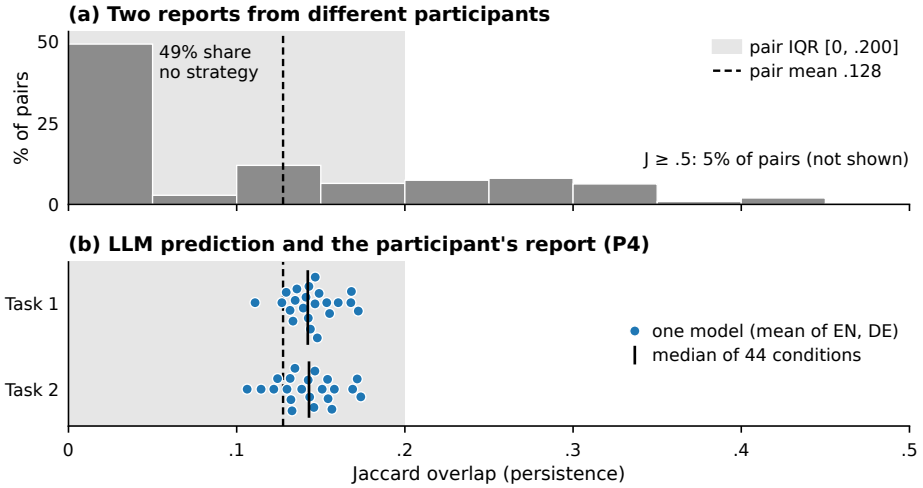

Figure 10: LLM Jaccard lies within the range of overlap between two different participants' reports. (a) Jaccard overlap between two persistence reports, over the 496,453 different-participant pairs among the 999 persistence reports that list at least one strategy (bins of width .05; the first bar is $J = 0$, no shared strategy). The grey band is the pair interquartile range $[0, .200]$ and the dashed line the pair mean (.128; median .083); the 5% of pairs with $J \geq .5$ lie outside the axis. (b) P4 Jaccard between each LLM's predicted strategies and the participant's report, one dot per model (mean of the English and German conditions); black lines mark the median of the 44 model–language conditions (Task 1 .142, Task 2 .143). Over the pairs in which at least one report lists a strategy, the pair mean is .100, below all 22 model means.

Table 39: Persistence Task 3 P4−P1 success-error movement.

Metric	Direction summarized	Total conditions	English	German	Holm conditions
absolute success error	lower error	+35/−9	+20/−2	+15/−7	+9/−0
signed success error	downward shift	downward 38/44	downward 19/22	downward 19/22	downward 32/44

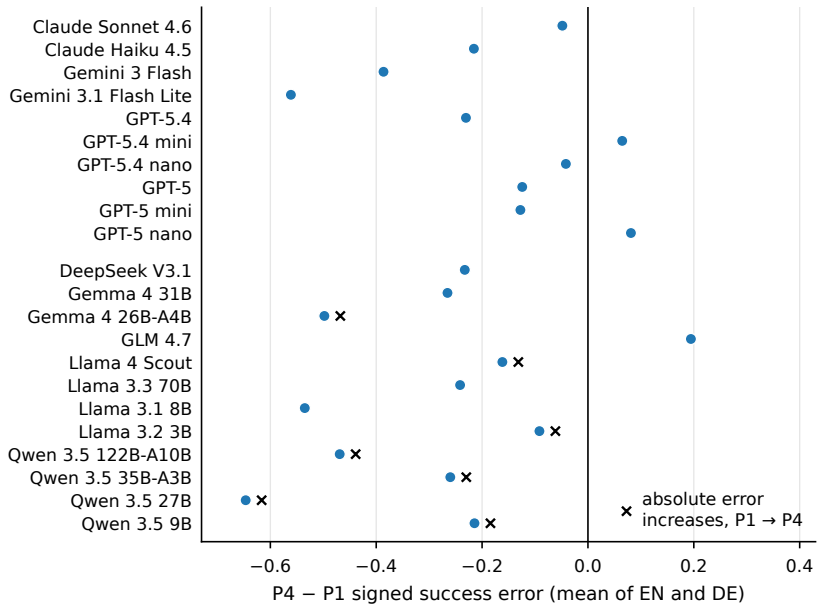

Figure 11: P4 lowers predicted success relative to P1 in 19 of 22 models (Task 3, persistence). Each dot is one model's P4−P1 change in signed success error (predicted minus reported success, 1–7 scale), averaged over the English and German conditions; negative values mean lower predicted success under P4. Rows follow the model order of Table 22 (closed models above the gap). × marks the seven models whose absolute success error is higher under P4 than under P1 (mean of English and German); all seven are open-weight and shift downward.

G Model-Level Evidence

Model-level rows evaluate whether the main patterns concentrate in a single model family or recur across families. The compact model-language tables document the systems that contribute to each pattern and the systems that depart from it.

G.1 Model-Level Results for Persistence

Table 40 gives each model's Task 1 values at P4, the numbers behind Figure 1(a, b), and Table 41 gives the values plotted in Figure 2.

The model-level breakdown separates two empirical facts. Several models reduce Task 3 absolute success error under P4 in one or both languages, and their P4−P1 changes in Tasks 1–2 remain close to zero. DeepSeek V3.1, GPT-5 mini, Gemini 3.1 Flash Lite, GPT-5.4, GLM 4.7, Llama 3.1 8B, and Llama 3.3 70B have mean Task 3 benefit above $+0.09$; their Task 1–2 movements range from -0.032 to $+0.012$ (Table 42). The pattern supports separate evaluation of success-score estimation and recovery of the reported strategies.

Tables 40 and 42 and Figure 10(b) constrain a simple model-size account. Gemma 4 31B carries the largest P4 Jaccard mean in both Task 1 and Task 2, while its smaller counterpart Gemma 4 26B-A4B moves away from participant reports across the three primary metrics. Closed-model rows split as well: Claude Sonnet 4.6 and Gemini 3 Flash are next behind Gemma 4 31B on raw P4 Jaccard, while GPT-5 mini and GPT-5.4 improve Task 3 with Holm significance in both languages and GPT-5.4 nano has small or mixed movements. The heterogeneity strengthens the process-level claim because the same limitation spans providers and scale tiers.

G.2 Family Aggregation

The family-level aggregation shows heterogeneity within families (Table 43). Gemma is the only family with negative medians on both Task 1 Jaccard movement and Task 3 error reduction; all

Table 40: Task 1 at P4 by model (persistence; participant-balanced means averaged over English and German). Same-size frequency: Jaccard of the held-out frequency ranking truncated at the model's own set size (grey dots in Figure 1(b)). For reference, the top-3 frequency rule reaches Jaccard .204, precision .330, recall .397, and F1 .300, and participants report 3.1 strategies per episode; recall is computed over reports that list at least one strategy.

Model	Jaccard	Precision	Recall	F1	Strategies per episode	Same-size frequency
Claude Sonnet 4.6	0.168	0.224	0.505	0.264	6.67	0.206
Claude Haiku 4.5	0.136	0.169	0.482	0.219	8.68	0.198
Gemini 3 Flash	0.168	0.197	0.596	0.262	9.22	0.199
Gemini 3.1 Flash Lite	0.147	0.179	0.543	0.235	8.87	0.201
GPT-5.4	0.155	0.191	0.513	0.244	8.69	0.202
GPT-5.4 mini	0.143	0.166	0.542	0.226	10.74	0.190
GPT-5.4 nano	0.133	0.149	0.635	0.218	13.19	0.178
GPT-5	0.140	0.148	0.800	0.228	16.68	0.160
GPT-5 mini	0.129	0.140	0.707	0.213	15.86	0.162
GPT-5 nano	0.127	0.156	0.462	0.202	9.53	0.195
DeepSeek V3.1	0.154	0.182	0.578	0.244	9.69	0.195
Gemma 4 31B	0.172	0.225	0.482	0.264	6.67	0.206
Gemma 4 26B-A4B	0.132	0.177	0.397	0.209	6.85	0.205
GLM 4.7	0.149	0.172	0.639	0.240	11.48	0.185
Llama 4 Scout	0.143	0.176	0.536	0.232	8.85	0.199
Llama 3.3 70B	0.147	0.205	0.384	0.230	5.81	0.205
Llama 3.1 8B	0.144	0.163	0.678	0.234	12.62	0.182
Llama 3.2 3B	0.111	0.156	0.344	0.183	6.46	0.201
Qwen 3.5 122B-A10B	0.148	0.192	0.449	0.234	7.17	0.202
Qwen 3.5 35B-A3B	0.160	0.208	0.472	0.250	7.02	0.200
Qwen 3.5 27B	0.141	0.177	0.493	0.228	8.46	0.200
Qwen 3.5 9B	0.135	0.181	0.398	0.214	7.00	0.203

four Qwen models, Llama 4 Scout, and Llama 3.2 3B have higher error at P4, so the split does not follow size. DeepSeek shows the largest family-median Task 3 movement (+0.146), followed by GLM (+0.101). The Knowledge rank median is highest in DeepSeek (0.204), Qwen (0.202), Claude (0.198), and GPT-5 (0.192); GLM and Llama trail at 0.085–0.091. The rank ordering is consistent with the Knowledge claim from Section 4.4.

Table 44 counts the favorable directions among six family-median contrasts per family: P4−P1 and P4−P3 on each of the three primary persistence metrics. No family is favorable on all six contrasts; DeepSeek and Gemini reach the highest count (5/6).

G.3 Task 3 Change Across Conflict Types

Task 3 P4−P1 absolute-success-error movement changes across continuing, starting, and resisting (Table 45). A model that moves favorably in persistence may move differently in inhibition. The rows constrain provider- or size-only accounts by showing conflict-type-dependent direction changes within the same model families.

Table 41: Values plotted in Figure 2, with the Task 3 errors (means of the English and German conditions; Tasks 1 and 3 persistence). ΔJ: Task 1 Jaccard at P4 minus P1. Error: participant-balanced absolute error of predicted success (1–7 scale). ΔB and ΔW: P4 minus P1 change in the between-person and within-person correlations of predicted with reported success. ρ: Task 4 Spearman correlation with observed MISCS Knowledge (K) and Regulation (R) scores.

Model	ΔJ	Error P1	Error P4	ΔB	ΔW	K ρ	R ρ
Claude Sonnet 4.6	+0.002	1.234	1.208	+0.056	−0.018	+0.196	+0.008
Claude Haiku 4.5	0.000	1.312	1.238	+0.074	+0.030	+0.201	−0.008
Gemini 3 Flash	+0.006	1.375	1.320	+0.070	+0.007	+0.182	−0.001
Gemini 3.1 Flash Lite	+0.005	1.387	1.271	+0.137	+0.051	+0.158	−0.030
GPT-5.4	+0.005	1.291	1.187	+0.159	+0.029	+0.163	−0.016
GPT-5.4 mini	+0.009	1.252	1.170	+0.098	+0.016	+0.181	+0.020
GPT-5.4 nano	−0.005	1.364	1.314	+0.050	−0.028	+0.177	+0.018
GPT-5	+0.004	1.386	1.298	+0.113	+0.033	+0.215	+0.031
GPT-5 mini	0.000	1.352	1.230	+0.124	−0.011	+0.192	+0.072
GPT-5 nano	+0.003	1.291	1.255	+0.120	+0.055	+0.090	−0.012
DeepSeek V3.1	+0.006	1.492	1.346	+0.166	+0.020	+0.204	−0.019
Gemma 4 31B	+0.004	1.364	1.301	+0.087	0.000	+0.212	+0.005
Gemma 4 26B-A4B	−0.025	1.310	1.450	+0.100	+0.017	+0.163	+0.024
GLM 4.7	−0.005	1.446	1.346	+0.159	+0.017	+0.091	−0.034
Llama 4 Scout	−0.007	1.408	1.415	+0.043	−0.004	+0.154	−0.022
Llama 3.3 70B	−0.004	1.411	1.319	+0.148	+0.009	+0.118	−0.039
Llama 3.1 8B	+0.008	1.464	1.364	+0.125	+0.003	+0.051	−0.024
Llama 3.2 3B	+0.003	1.234	1.259	+0.096	−0.026	+0.043	+0.012
Qwen 3.5 122B-A10B	+0.010	1.334	1.345	+0.107	−0.005	+0.221	+0.008
Qwen 3.5 35B-A3B	+0.009	1.349	1.360	+0.152	+0.051	+0.145	−0.016
Qwen 3.5 27B	0.000	1.300	1.312	+0.128	+0.034	+0.196	+0.045
Qwen 3.5 9B	−0.011	1.245	1.275	+0.165	−0.004	+0.208	−0.028

Table 42: Persistence P4−P1 movement by model and language. EN and DE denote English and German prompt conditions; stars mark Holm-significant conditions in the corresponding family. Positive values mean higher Jaccard overlap for Tasks 1–2 or lower absolute success error for Task 3; signed success error is a calibration shift.

	Task 1 Jaccard		Task 2 Jaccard		Task 3 absolute error		Task 3 signed error	
Model	EN	DE	EN	DE	EN	DE	EN	DE
Claude Sonnet 4.6	+0.006	−0.001	−0.001	+0.002	+0.021	+0.030	−0.089	−0.008
Claude Haiku 4.5	+0.001	−0.002	+0.004	+0.001	+0.074	+0.075	−0.169*	−0.262*
Gemini 3 Flash	+0.005	+0.006	+0.003	+0.006	+0.059	+0.051	−0.283*	−0.490*
Gemini 3.1 Flash Lite	+0.010	−0.000	+0.010	+0.008	+0.179*	+0.052	−0.399*	−0.723*
GPT-5.4	−0.001	+0.012	−0.004	+0.007	+0.114*	+0.094*	−0.154*	−0.307*
GPT-5.4 mini	+0.008	+0.010	+0.006	+0.016*	+0.083	+0.082	+0.002	+0.128*
GPT-5.4 nano	−0.004	−0.006	−0.006	−0.006	+0.093*	+0.007	−0.064	−0.019
GPT-5	+0.001	+0.006*	+0.001	+0.007*	+0.013	+0.164*	−0.200*	−0.048
GPT-5 mini	−0.001	−0.000	+0.002	−0.003	+0.118*	+0.128*	−0.136*	−0.119*
GPT-5 nano	+0.013*	−0.006	+0.005	−0.001	+0.084	−0.012	+0.191*	−0.028
DeepSeek V3.1	+0.011	+0.001	+0.012	−0.008	+0.172*	+0.120	−0.283*	−0.182*
Gemma 4 31B	+0.004	+0.004	+0.018	+0.025*	+0.042	+0.084	−0.333*	−0.198*
Gemma 4 26B-A4B	−0.029*	−0.021*	−0.018*	−0.014	−0.131	−0.147	−0.547*	−0.449*
GLM 4.7	−0.004	−0.005	+0.003	−0.005	+0.059	+0.143	+0.049	+0.340*
Llama 4 Scout	−0.006	−0.008	−0.022*	−0.014*	+0.057	−0.071	−0.090*	−0.234*
Llama 3.3 70B	−0.001	−0.007	+0.010	+0.006	+0.132*	+0.053	−0.204*	−0.279*
Llama 3.1 8B	+0.012	+0.004	−0.032*	−0.016*	+0.136	+0.062	−0.625*	−0.444*
Llama 3.2 3B	+0.006	+0.001	+0.013*	−0.008	+0.006	−0.057	−0.188*	+0.005
Qwen 3.5 122B-A10B	+0.012*	+0.007	+0.003	+0.013*	+0.000	−0.022	−0.475*	−0.463*
Qwen 3.5 35B-A3B	+0.010	+0.009	+0.012	−0.003	+0.015	−0.036	−0.254*	−0.266*
Qwen 3.5 27B	+0.001	−0.001	+0.016*	+0.011	+0.029	−0.054	−0.525*	−0.768*
Qwen 3.5 9B	+0.001	−0.024*	+0.016*	−0.020*	−0.082	+0.020	−0.340*	−0.088*

Table 43: Persistence movement aggregated by model family. Entries are medians over each family's models (English and German averaged per model first) of the P4−P1 change in Task 1 Jaccard and Task 3 absolute success error, and of Task 4 Knowledge ρ. The Task 1 column carries the action-overlap pattern, which Task 2 follows in the main persistence summary. Positive values indicate higher Jaccard, lower absolute success error, or higher Knowledge ρ. GPT-5.4 and GPT-5 are listed as separate families.

Family	Models	Task 1 Jaccard overlap	Task 3 absolute success error	Task 4 Knowledge ρ
Claude	2	+0.001	+0.050	0.198
Gemini	2	+0.005	+0.085	0.170
GPT-5.4	3	+0.005	+0.082	0.177
GPT-5	3	+0.003	+0.088	0.192
DeepSeek	1	+0.006	+0.146	0.204
Gemma	2	−0.010	−0.038	0.188
GLM	1	−0.005	+0.101	0.091
Llama	4	−0.001	+0.043	0.085
Qwen	4	+0.005	−0.012	0.202

Table 44: Persistence prompt-block movement summarized by model family. Each count is over six family-median contrasts (three metrics × P4−P1 and P4−P3; the metrics are Task 1 Jaccard overlap, Task 2 Jaccard overlap, and Task 3 absolute success error); a contrast is favorable when the family median indicates higher Jaccard or lower absolute success error. GPT-5.4 and GPT-5 are listed as separate families.

Family	Models	Favorable contrasts
Claude	2	4/6
Gemini	2	5/6
GPT-5.4	3	4/6
GPT-5	3	3/6
DeepSeek	1	5/6
Gemma	2	3/6
GLM	1	4/6
Llama	4	3/6
Qwen	4	3/6

Table 45: Task 3 absolute-success-error P4−P1 movement by conflict type, model, and language. EN and DE denote English and German prompt conditions; stars mark Holm-significant conditions. Positive values mean lower absolute success error under P4 than P1.

Model	Persistence	Initiation	Inhibition
Claude Sonnet 4.6	EN +0.021 / DE +0.030	EN −0.005 / DE −0.019	EN −0.125 / DE −0.056
Claude Haiku 4.5	EN +0.074 / DE +0.075	EN +0.031 / DE +0.033	EN −0.022 / DE −0.005
Gemini 3 Flash	EN +0.059 / DE +0.051	EN −0.092 / DE −0.193	EN −0.140 / DE −0.216
Gemini 3.1 Flash Lite	EN +0.179* / DE +0.052	EN +0.034 / DE −0.112	EN −0.101 / DE −0.119
GPT-5.4	EN +0.114* / DE +0.094*	EN +0.034 / DE +0.064	EN −0.016 / DE +0.015
GPT-5.4 mini	EN +0.083 / DE +0.082	EN +0.028 / DE +0.017	EN +0.008 / DE −0.032
GPT-5.4 nano	EN +0.093* / DE +0.007	EN +0.055 / DE +0.030	EN +0.043 / DE −0.005
GPT-5	EN +0.013 / DE +0.164*	EN +0.010 / DE +0.026	EN −0.030 / DE −0.039
GPT-5 mini	EN +0.118* / DE +0.128*	EN +0.019 / DE +0.044	EN −0.042 / DE +0.004
GPT-5 nano	EN +0.084 / DE −0.012	EN +0.010 / DE +0.025	EN +0.003 / DE +0.036
DeepSeek V3.1	EN +0.172* / DE +0.120	EN −0.041 / DE −0.056	EN −0.107 / DE −0.069
Gemma 4 31B	EN +0.042 / DE +0.084	EN −0.032 / DE −0.054	EN −0.054 / DE −0.086
Gemma 4 26B-A4B	EN −0.131 / DE −0.147	EN −0.140* / DE −0.199*	EN −0.190 / DE −0.222*
GLM 4.7	EN +0.059 / DE +0.143	EN −0.066 / DE +0.010	EN −0.013 / DE −0.029
Llama 4 Scout	EN +0.057 / DE −0.071	EN −0.090* / DE −0.093*	EN −0.081 / DE −0.064
Llama 3.3 70B	EN +0.132* / DE +0.053	EN +0.047 / DE +0.034	EN −0.028 / DE +0.018
Llama 3.1 8B	EN +0.136 / DE +0.062	EN −0.022 / DE −0.080	EN −0.013 / DE −0.022
Llama 3.2 3B	EN +0.006 / DE −0.057	EN −0.068 / DE −0.086*	EN −0.093 / DE −0.149
Qwen 3.5 122B-A10B	EN +0.000 / DE −0.022	EN −0.042 / DE −0.062	EN −0.218 / DE −0.206
Qwen 3.5 35B-A3B	EN +0.015 / DE −0.036	EN −0.069 / DE −0.086	EN −0.116 / DE −0.237*
Qwen 3.5 27B	EN +0.029 / DE −0.054	EN −0.052 / DE −0.115	EN −0.119 / DE −0.175
Qwen 3.5 9B	EN −0.082 / DE +0.020	EN −0.061 / DE −0.036	EN −0.062 / DE −0.068

H Other Conflict Types

Starting and resisting episodes serve as secondary checks for the persistence claim. The same prompt blocks and tasks are evaluated under different conflict types: continuing, starting, and resisting an action. These are the persistence, initiation, and inhibition conflict types.

H.1 Summary Across Conflict Types

Conflict type changes the target. The same prediction task has a different psychological burden when the participant is continuing, starting, or resisting an action. Continuing concerns maintenance of an ongoing activity. Starting concerns movement from intention to action. Resisting concerns suppression or avoidance of an unwanted action. These conflict types share the label set and prompt blocks. Their success priors and strategy distributions differ.

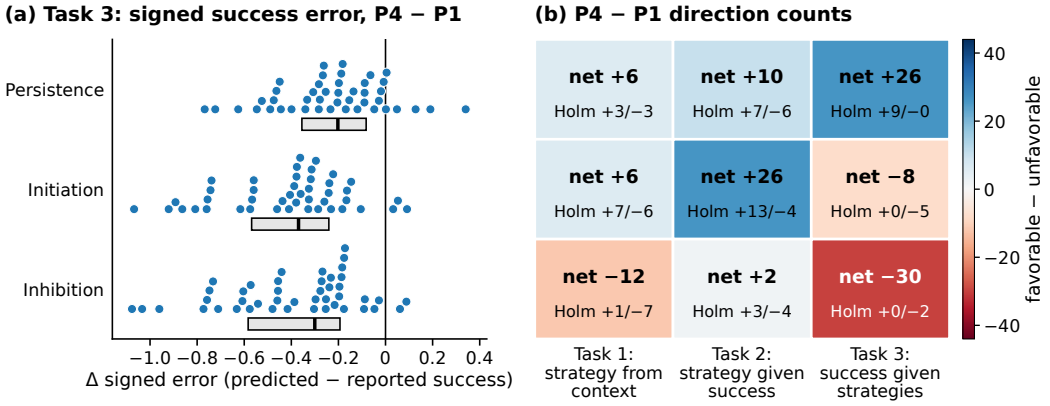

Figure 12: P4 lowers predicted success in all three conflict types; this reduces Task 3 error for persistence but increases it for initiation and inhibition. Conflict types: persistence (continuing an aversive activity), initiation (starting one), and inhibition (resisting a temptation). (a) P4−P1 change in the signed success error (predicted minus reported success, 1–7 scale) of each of the 44 model–language conditions per conflict type (dots); boxes span the interquartile range and the thick line marks the median. The change is negative in 38 (persistence), 41 (initiation), and 42 (inhibition) of 44 conditions. (b) P4−P1 direction counts for Tasks 1–3. "net" is the number of conditions with a favorable change minus the number with an unfavorable change (color), where favorable means higher Jaccard in Tasks 1–2 and lower absolute success error in Task 3; "Holm" gives the Holm-significant favorable/unfavorable conditions. The absolute Task 3 error falls in 35 of 44 persistence conditions but rises in 26 of 44 initiation and 37 of 44 inhibition conditions. Exact counts and medians are in Tables 46 and 47.

Conflict-type-specific change. The P4 pattern differs by conflict type (Figures 12 and 13). Persistence has its clearest P4-direction pattern in Task 3, initiation in Task 2, and inhibition often moves away from participant success reports in Task 3. Participant-report distance shows the same pattern as in persistence (Table 48): across the three conflict types, the median distance is 0.851–0.858 for Tasks 1–2 and 1.306–1.421 for Task 3, so relative movement leaves the predictions far from the participant reports.

H.2 Initiation

Starting favors conditional prediction. Initiation concentrates the strongest favorable strategy movement in Task 2. P4 improves Task 2 Jaccard overlap in 35/44 model-language conditions, with 13 favorable Holm-significant conditions and a median signed movement of $+0.008$ (Table 50). Task 1 remains weaker at 25/44 favorable conditions and a median movement of $+0.003$. The P4 persona blocks thus improve strategy prediction most clearly when the prompt also supplies the reported success and the episode concerns starting an activity. The predicted sets keep the persistence pattern:

Table 46: P4−P1 movement by conflict type and task. Positive signed values mean higher Jaccard for Tasks 1–2 and lower absolute success error for Task 3; counts are fixed model-language conditions.

Conflict	Task	Metric	Conditions	Holm	Med. signed Δ [IQR]	Med. signed r_{rb} [IQR]
Persistence	Task 1	Jaccard overlap	+25/−19	+3/−3	+0.001 [−0.002, +0.006]	+0.040 [−0.103, +0.134]
Persistence	Task 2	Jaccard overlap	+27/−17	+7/−6	+0.003 [−0.005, +0.010]	+0.043 [−0.092, +0.162]
Persistence	Task 3: strategy to success	absolute success error	+35/−9	+9/−0	+0.055 [+0.007, +0.093]	+0.101 [+0.013, +0.245]
Initiation	Task 1	Jaccard overlap	+25/−19	+7/−6	+0.003 [−0.003, +0.007]	+0.024 [−0.105, +0.129]
Initiation	Task 2	Jaccard overlap	+35/−9	+13/−4	+0.008 [+0.003, +0.013]	+0.155 [+0.014, +0.230]
Initiation	Task 3: strategy to success	absolute success error	+18/−26	+0/−5	−0.034 [−0.072, +0.026]	−0.057 [−0.158, +0.086]
Inhibition	Task 1	Jaccard overlap	+16/−28	+1/−7	−0.002 [−0.010, +0.003]	−0.101 [−0.197, +0.032]
Inhibition	Task 2	Jaccard overlap	+23/−21	+3/−4	+0.001 [−0.006, +0.009]	+0.000 [−0.120, +0.158]
Inhibition	Task 3: strategy to success	absolute success error	+7/−37	+0/−2	−0.059 [−0.119, −0.016]	−0.129 [−0.205, −0.019]

Table 47: Task 3 signed-success-error P4−P1 shift by conflict type. Downward conditions mean P4 lowers predicted success relative to P1.

Conflict	Downward conditions	Upward conditions	Tie conditions	Holm downward	Holm upward	Median signed Δ [IQR]
Persistence	38/44	6/44	0/44	32/44	3/44	−0.202 [−0.355, −0.082]
Initiation	41/44	3/44	0/44	41/44	2/44	−0.369 [−0.569, −0.241]
Inhibition	42/44	2/44	0/44	39/44	1/44	−0.300 [−0.583, −0.193]

Table 48: P4 participant-report distance by conflict type: 1 − Jaccard for Tasks 1–2 and absolute success error (1–7 scale) for Task 3; medians and interquartile ranges over the 44 model-language conditions of each conflict type.

Conflict	Task	Primary metric	Median distance	Distance IQR
Persistence	Task 1	Jaccard overlap	0.858	[0.845, 0.864]
Persistence	Task 2	Jaccard overlap	0.857	[0.846, 0.872]
Persistence	Task 3: strategy to success	Absolute success error	1.306	[1.243, 1.348]
Initiation	Task 1	Jaccard overlap	0.853	[0.844, 0.859]
Initiation	Task 2	Jaccard overlap	0.851	[0.837, 0.864]
Initiation	Task 3: strategy to success	Absolute success error	1.321	[1.256, 1.387]
Inhibition	Task 1	Jaccard overlap	0.856	[0.851, 0.862]
Inhibition	Task 2	Jaccard overlap	0.857	[0.846, 0.870]
Inhibition	Task 3: strategy to success	Absolute success error	1.421	[1.384, 1.532]

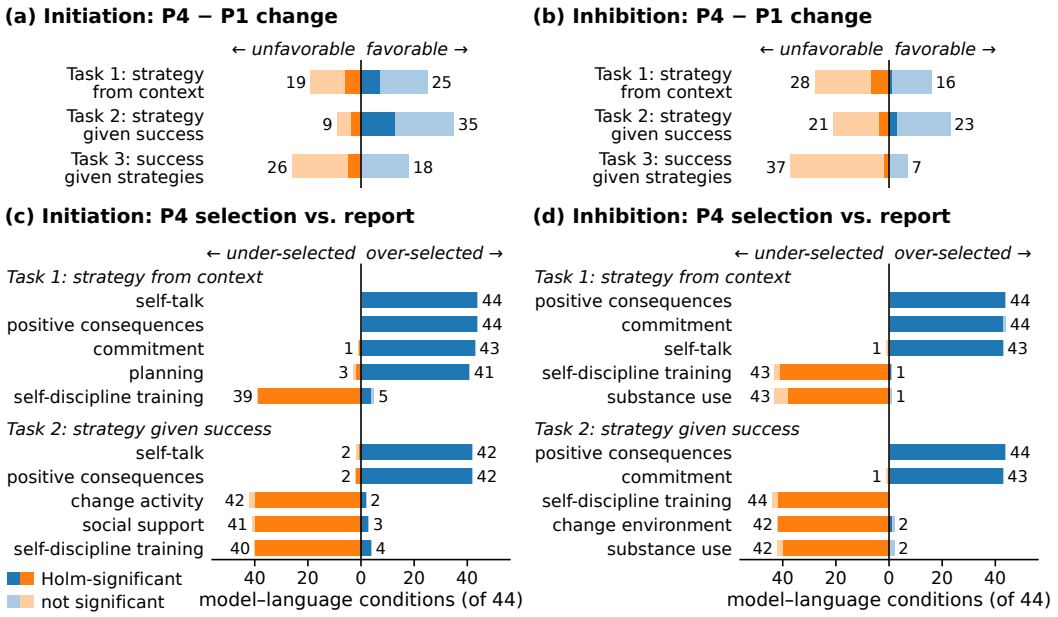

Figure 13: Initiation and inhibition summaries. P4 improves initiation Task 2 in 35 of 44 conditions, inhibition tasks split evenly or move unfavorably, and P4 keeps the persistence label skew in both conflict types. (a, b) Number of the 44 model–language conditions whose P4−P1 change is favorable (right; higher Jaccard in Tasks 1–2, lower absolute success error in Task 3) or unfavorable (left). (c, d) For the labels of Tables 51 and 54, number of conditions in which the P4 selection rate is above (over-selected, right) or below (under-selected, left) the participants' report rate; the difference of the two bars is the over- minus under-selection count. Darker segments: Holm-significant conditions. Holm counts are episode-level McNemar tests; the participant-level check does not confirm the self-discipline training direction in persistence or inhibition (Appendix B.4). Initiation: starting an aversive activity; inhibition: resisting a temptation.

Table 49: Initiation P4 strategy-set size (labels per episode; 44 model–language conditions).

Task	Participant mean	P4 median	P4 range	Above participant mean	English	German
Task 1	3.338	8.777	5.554–17.398	44/44	22/22	22/22
Task 2	3.338	6.395	4.207–13.629	44/44	22/22	22/22

every P4 condition predicts more strategies than participants report (Table 49), and self-talk, positive consequences, commitment, and planning are over-selected in 41–44 of 44 conditions (Table 51).

Table 50: Initiation named prompt-block contrasts. Counts are 44 model-language conditions per row; Holm conditions are corrected within the stated metric family.

Task	Metric	Contrast	Conditions	Holm	Med. signed Δ [IQR]	Med. signed r_{rb} [IQR]
Task 1	Jaccard overlap	P2−P1	+27/−17	+6/−7	+0.002 [−0.003, +0.005]	+0.031 [−0.117, +0.119]
Task 1	Jaccard overlap	P3−P2	+27/−17	+4/−1	+0.001 [−0.002, +0.004]	+0.017 [−0.087, +0.112]
Task 1	Jaccard overlap	P4−P3	+19/−25	+5/−5	−0.001 [−0.003, +0.002]	+0.022 [−0.099, +0.101]
Task 1	Jaccard overlap	P4−P1	+25/−19	+7/−6	+0.003 [−0.003, +0.007]	+0.024 [−0.105, +0.129]
Task 2	Jaccard overlap	P2−P1	+32/−12	+7/−4	+0.003 [−0.001, +0.006]	+0.060 [−0.040, +0.143]
Task 2	Jaccard overlap	P3−P2	+30/−14	+6/−2	+0.002 [−0.001, +0.006]	+0.055 [−0.059, +0.130]
Task 2	Jaccard overlap	P4−P3	+30/−14	+4/−1	+0.001 [−0.001, +0.005]	+0.077 [−0.002, +0.124]
Task 2	Jaccard overlap	P4−P1	+35/−9	+13/−4	+0.008 [+0.003, +0.013]	+0.155 [+0.014, +0.230]
Task 3: strategy to success	absolute success error	P2−P1	+29/−15	+2/−1	+0.012 [−0.010, +0.024]	+0.045 [−0.035, +0.141]
Task 3: strategy to success	absolute success error	P3−P2	+14/−30	+0/−4	−0.015 [−0.036, +0.006]	−0.068 [−0.158, +0.031]
Task 3: strategy to success	absolute success error	P4−P3	+16/−28	+2/−7	−0.012 [−0.031, +0.005]	−0.077 [−0.248, +0.020]
Task 3: strategy to success	absolute success error	P4−P1	+18/−26	+0/−5	−0.034 [−0.072, +0.026]	−0.057 [−0.158, +0.086]

Narrow initiation gain. The initiation gain appears in Task 2, which conditions strategy prediction on the reported success. Task 3 moves in the unfavorable direction in 26/44 conditions, including 5 unfavorable Holm-significant conditions, and the median absolute-success-error movement is -0.034 (Table 46). Starting an action also changes strategy-situation fit; the evaluation target remains the participant-reported strategy set for that episode.

Table 51: Initiation P4 strategy-label direction tests across 44 model-language conditions per row.

Task	Strategy label	Direction	Direction conditions	Holm conditions	Median rate Δ [IQR]
Task 1	Self-talk	over	44/44	44/44	+0.903 [+0.761, +0.930]
Task 1	Positive consequences	over	44/44	44/44	+0.498 [+0.465, +0.501]
Task 1	Commitment	over	43/44	43/44	+0.711 [+0.564, +0.837]
Task 1	Planning	over	41/44	41/44	+0.799 [+0.513, +0.828]
Task 1	Self-discipline training	under	39/44	39/44	−0.017 [−0.017, −0.017]
Task 2	Self-talk	over	42/44	42/44	+0.780 [+0.537, +0.914]
Task 2	Positive consequences	over	42/44	42/44	+0.428 [+0.362, +0.477]
Task 2	Change activity	under	42/44	40/44	−0.075 [−0.075, −0.074]
Task 2	Social support	under	41/44	40/44	−0.053 [−0.053, −0.052]
Task 2	Self-discipline training	under	40/44	40/44	−0.017 [−0.017, −0.017]

H.3 INHIBITION

Table 52: Inhibition P4 strategy-set size (labels per episode; 44 model–language conditions).

Task	Participant mean	P4 median	P4 range	Above participant mean	English	German
Task 1	2.826	7.427	4.482–15.811	44/44	22/22	22/22
Task 2	2.826	5.787	3.826–11.089	44/44	22/22	22/22

Resisting changes success. Inhibition shows the clearest error increase. P4 lowers predicted success in 42/44 conditions, 39 after Holm correction. Task 3 absolute-success-error movement is unfavorable in 37/44 conditions, including 2 unfavorable Holm-significant conditions (Tables 46 and 47; block contrasts in Table 53). Lower predicted success can reduce persistence error when P1 predicts higher success than participants report; in resisting episodes, the same directional shift often moves away from the participant's success appraisal.

Familiar labels miss reported resisting tactics. In inhibition, every P4 condition predicts more strategies than participants report (Table 52). P4 over-selects positive consequences, commitment, and self-talk (43–44 of 44 conditions) and under-selects self-discipline training, substance use, and changing the environment (42–44 of 44; Table 54), the self-discipline direction in episode-level tests only (Appendix B.4). The labels that P4 over-selects in persistence thus recur when the participant resists a temptation.

Table 53: Inhibition named prompt-block contrasts. Counts are 44 model-language conditions per row; Holm conditions are corrected within the stated metric family.

Task	Metric	Contrast	Conditions	Holm	Med. signed Δ [IQR]	Med. signed r_{rb} [IQR]
Task 1	Jaccard overlap	P2−P1	+11/−33	+1/−3	−0.004 [−0.009, −0.001]	−0.125 [−0.208, −0.031]
Task 1	Jaccard overlap	P3−P2	+20/−24	+1/−3	−0.001 [−0.006, +0.003]	−0.010 [−0.146, +0.071]
Task 1	Jaccard overlap	P4−P3	+25/−19	+4/−0	+0.001 [−0.002, +0.004]	+0.030 [−0.031, +0.133]
Task 1	Jaccard overlap	P4−P1	+16/−28	+1/−7	−0.002 [−0.010, +0.003]	−0.101 [−0.197, +0.032]
Task 2	Jaccard overlap	P2−P1	+25/−19	+1/−3	+0.001 [−0.003, +0.007]	−0.007 [−0.089, +0.103]
Task 2	Jaccard overlap	P3−P2	+21/−23	+5/−3	−0.000 [−0.003, +0.005]	−0.005 [−0.076, +0.115]
Task 2	Jaccard overlap	P4−P3	+20/−24	+1/−0	−0.000 [−0.005, +0.003]	−0.002 [−0.120, +0.085]
Task 2	Jaccard overlap	P4−P1	+23/−21	+3/−4	+0.001 [−0.006, +0.009]	+0.000 [−0.120, +0.158]
Task 3: strategy to success	absolute success error	P2−P1	+15/−29	+0/−1	−0.022 [−0.046, +0.007]	−0.059 [−0.186, +0.035]
Task 3: strategy to success	absolute success error	P3−P2	+9/−35	+0/−3	−0.031 [−0.058, −0.003]	−0.109 [−0.183, −0.018]
Task 3: strategy to success	absolute success error	P4−P3	+17/−27	+1/−2	−0.010 [−0.030, +0.009]	−0.057 [−0.169, +0.065]
Task 3: strategy to success	absolute success error	P4−P1	+7/−37	+0/−2	−0.059 [−0.119, −0.016]	−0.129 [−0.205, −0.019]

Table 54: Inhibition P4 strategy-label direction tests across 44 model-language conditions per row. Holm counts are episode-level McNemar tests; the participant-level check does not confirm the self-discipline training direction in persistence or inhibition (Appendix B.4).

Task	Strategy label	Direction	Direction conditions	Holm conditions	Median rate Δ [IQR]
Task 1	Positive consequences	over	44/44	44/44	+0.485 [+0.425, +0.513]
Task 1	Commitment	over	44/44	43/44	+0.816 [+0.733, +0.884]
Task 1	Self-talk	over	43/44	43/44	+0.885 [+0.619, +0.943]
Task 1	Self-discipline training	under	43/44	41/44	−0.023 [−0.023, −0.023]
Task 1	Substance use	under	43/44	38/44	−0.076 [−0.076, −0.064]
Task 2	Positive consequences	over	44/44	44/44	+0.384 [+0.338, +0.451]
Task 2	Commitment	over	43/44	43/44	+0.698 [+0.594, +0.785]
Task 2	Self-discipline training	under	44/44	42/44	−0.023 [−0.023, −0.023]
Task 2	Change environment	under	42/44	42/44	−0.220 [−0.222, −0.209]
Task 2	Substance use	under	42/44	40/44	−0.075 [−0.076, −0.070]

I TASK 4 AUDITS

Task 4 predicts the MISCS subscale scores from conflict histories. With the MISCS item responses withheld, the models' ratings rank Knowledge weakly, an association that mean success, strategies per episode, and repertoire size account for (Appendix A.7), and their Regulation associations stay near zero, although a ridge regression on the same histories ranks Regulation at $\rho = .158$.

I.1 CONSTRUCT DEFINITION

The construct definition separates Knowledge and Regulation before the rank tests are interpreted. Knowledge includes awareness, remembered conflicts, strengths and weaknesses, strategy knowledge, and strategy-situation fit. Regulation includes planning, monitoring, and evaluation. The rank tests are therefore interpreted separately for Knowledge and Regulation.

I.2 RANK RESULTS WITH EQUIVALENCE TESTS

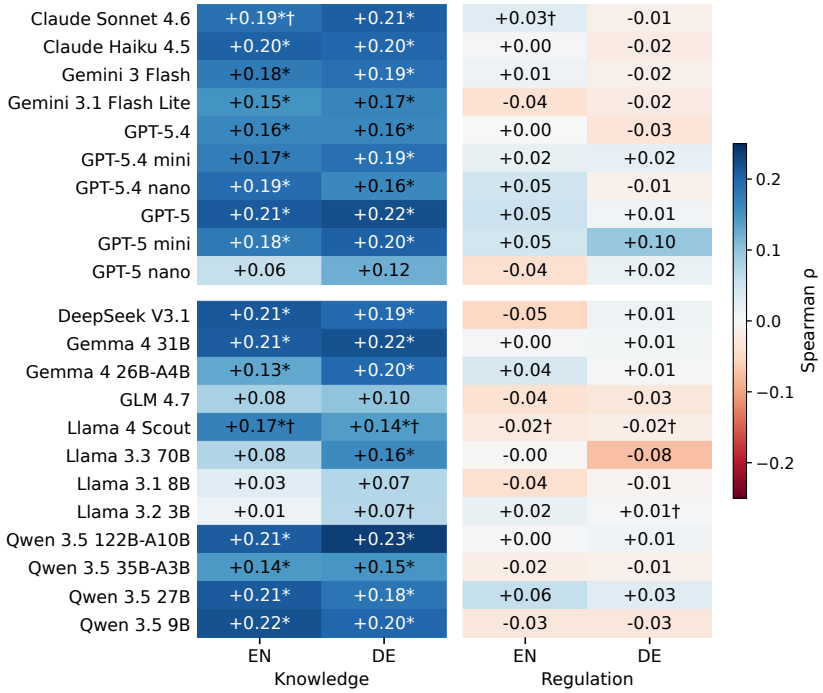

Figure 14: LLM predictions from conflict histories rank MISCS Knowledge in most model–language conditions, and Regulation associations stay near zero (Task 4). Cells give the Spearman ρ between predicted and observed subscale scores for English (EN) and German (DE) prompts. $*$: Holm-significant within subscale (Knowledge 35 of 44 conditions, median $\rho = .178$; Regulation 0 of 44, median .003, all within $\pm.10$; the largest, .096, rounds to .10). †: aggregate-score condition, whose parsed item-level trace covers a subset of the rank-summary participants (Table 60). Rows follow the model order of Table 22; Figure 2(c, d) shows the means of the two languages.

The model-level rank breakdown (Table 56, Figure 14) is the full counterpart to Figure 2. Knowledge-score ordering is positive and Holm-significant in most of the model panel; Regulation remains close to zero in the corrected rank tests.

The equivalence test (Table 57) assesses whether the data support population rank associations within ± 0.10 on the Fisher-z scale ($|\rho| < .0997$), a margin chosen for this study. $|\rho| < 0.10$ holds for all 44 Regulation conditions and for 7/44 Knowledge conditions, and 25/44 Regulation conditions formally pass TOST equivalence. The JZS Bayes factor gives a median $\mathrm{BF}_{01} = 17.6$ for Regulation, above 3 in 43 of 44 conditions, and a median $\mathrm{BF}_{01} = 0.012$ for Knowledge, below

Table 55: Task 4 MISCS prediction by 44 model-language conditions. Rank rows report Spearman association with participant scores; bias and MAE rows summarize score-level calibration. Complete-item-trace rows use the 40 conditions with complete parsed item-level coverage, which trace each predicted subscale score to its parsed items.

Target / diagnostic	Statistic	Conditions	English	German	Median [IQR]	Range / calibration
Knowledge rank	Spearman ρ	Holm rank 35/44	17/22	18/22	+0.178 [+0.142, +0.204]	0.015–0.235
Knowledge signed bias	positive bias	positive 42/44	21/22	21/22	+0.617 [+0.480, +0.856]	MAE median 0.7452
Knowledge trace rank	Spearman ρ	positive 40/40; Holm 32/40	positive 20/20; Holm 15/20	positive 20/20; Holm 17/20	+0.181 [+0.146, +0.205]	0.015–0.235
Regulation rank	Spearman ρ	Holm rank 0/44	0/22	0/22	+0.003 [−0.022, +0.017]	−0.075–0.096
Regulation signed bias	positive bias	positive 29/44	14/22	15/22	+0.211 [−0.197, +0.392]	MAE median 0.7448
Regulation trace rank	Spearman ρ	positive 21/40; Holm 0/40	positive 12/20; Holm 0/20	positive 9/20; Holm 0/20	+0.003 [−0.023, +0.017]	−0.075–0.096

Table 56: Task 4 MISCS rank associations by model and language. EN and DE denote English and German prompt conditions. Marked rank values are Holm-significant; daggers identify aggregate-score conditions, whose parsed item-level trace covers a subset of the ranked participants.

Model	Knowledge ρ	Regulation ρ
Claude Sonnet 4.6	EN +0.186*† / DE +0.206*	EN +0.030† / DE −0.015
Claude Haiku 4.5	EN +0.204* / DE +0.198*	EN +0.004 / DE −0.021
Gemini 3 Flash	EN +0.176* / DE +0.188*	EN +0.013 / DE −0.016
Gemini 3.1 Flash Lite	EN +0.150* / DE +0.166*	EN −0.039 / DE −0.020
GPT-5.4	EN +0.163* / DE +0.162*	EN +0.001 / DE −0.033
GPT-5.4 mini	EN +0.173* / DE +0.188*	EN +0.019 / DE +0.022
GPT-5.4 nano	EN +0.192* / DE +0.162*	EN +0.050 / DE −0.013
GPT-5	EN +0.210* / DE +0.220*	EN +0.054 / DE +0.008
GPT-5 mini	EN +0.179* / DE +0.205*	EN +0.048 / DE +0.096
GPT-5 nano	EN +0.059 / DE +0.122	EN −0.040 / DE +0.016
DeepSeek V3.1	EN +0.213* / DE +0.195*	EN −0.051 / DE +0.013
Gemma 4 31B	EN +0.208* / DE +0.217*	EN +0.004 / DE +0.006
Gemma 4 26B-A4B	EN +0.130* / DE +0.197*	EN +0.042 / DE +0.005
GLM 4.7	EN +0.080 / DE +0.101	EN −0.039 / DE −0.029
Llama 4 Scout	EN +0.168*† / DE +0.141*†	EN −0.024† / DE −0.019†
Llama 3.3 70B	EN +0.076 / DE +0.160*	EN −0.003 / DE −0.075
Llama 3.1 8B	EN +0.030 / DE +0.073	EN −0.041 / DE −0.007
Llama 3.2 3B	EN +0.015 / DE +0.071†	EN +0.018 / DE +0.006†
Qwen 3.5 122B-A10B	EN +0.207* / DE +0.235*	EN +0.005 / DE +0.010
Qwen 3.5 35B-A3B	EN +0.143* / DE +0.147*	EN −0.019 / DE −0.013
Qwen 3.5 27B	EN +0.211* / DE +0.182*	EN +0.060 / DE +0.029
Qwen 3.5 9B	EN +0.218* / DE +0.197*	EN −0.028 / DE −0.028

Table 57: Task 4 equivalence and Bayes-factor diagnostics for Knowledge and Regulation rank associations. TOST uses a Fisher-z margin of ± 0.10 at $\alpha = 0.05$. BF_{01} is the JZS Bayes factor for zero rank association (Cauchy scale $\sqrt{2}/2$).

Subscale	Conditions	Median ρ	TOST equivalent (± 0.10)	$\lvert\rho\rvert < 0.10$ conditions	Median BF_{01}	$BF_{01} > 3$ conditions
Knowledge	44	0.178	1/44	7/44	0.012	7/44
Regulation	44	0.003	25/44	44/44	17.619	43/44

1/3 in 34 of 44 conditions. The two subscales therefore contrast under the same margin and test: Knowledge is weakly rank-ordered (Holm-significant in 35 of 44 conditions), whereas Regulation remains inside the tested band, has no Holm-significant condition, and has Bayes factors favoring no rank association.

The Cronbach α values for the participant MISCS-12 subscales ($\alpha = 0.719$ for Knowledge, $\alpha = 0.786$ for Regulation; Table 27) are moderate for six-item subscales and set the measurement-error attenuation of the Task 4 rank associations. Task 4 is a subscale-level rank audit. Two implications follow. First, the observed Knowledge ρ is attenuated by measurement error; under classical test theory the attenuation factor is $\sqrt{\alpha} \approx 0.85$ and the disattenuation correction multiplier is $1/\sqrt{\alpha} \approx 1.18$, giving an outcome-reliability-only corrected Knowledge median near $\rho \approx 0.21$. Second, the same correction with the Regulation α applied to the equivalence margin gives $0.10/\sqrt{0.786} \approx 0.113$: an observed $\lvert\rho\rvert < 0.10$ corresponds to an outcome-reliability-corrected $\lvert\rho\rvert < 0.113$ (ticks in Figure 9(b)). The Regulation result is therefore a failure of the models to rank participants whose histories carry Regulation information (ridge $\rho = .158$; Appendix A.7), with all 44 conditions inside the $\pm .10$ band.

The complete-item-trace sensitivity in Table 55 uses the 40 model-language conditions with complete parsed-item coverage, which trace each predicted subscale score to its parsed items; the other four conditions are the aggregate-score conditions of Table 60. The Knowledge pattern remains positive in all 40 retained conditions and Holm-significant in 32. Regulation has no Holm-significant retained condition. The sensitivity supports the headline split under the stricter provenance subset.

Table 58: Task 4 participant denominators. Rank rows report the participants in each Spearman condition; HC3 rows report raw, analyzed, and excluded participants for the history-length adjusted regressions.

Target	Analysis	Conditions	Raw N	Analyzed N	Excluded N
Knowledge	Spearman rank	44	466–467	466–467	0
Regulation	Spearman rank	44	466–467	466–467	0
Knowledge	Absolute score error HC3	44	466–467	447–448	19
Knowledge	Signed score error HC3	44	466–467	447–448	19
Regulation	Absolute score error HC3	44	466–467	447–448	19
Regulation	Signed score error HC3	44	466–467	447–448	19

Rank association and score-error diagnostics use different participant denominators (Table 58). Spearman analyses use 466–467 participants per model-language condition. The HC3 diagnostics use 447–448 analyzed participants after covariate or outcome exclusions. Rank-test N can exceed the complete-six-item reliability N (464 for Knowledge, 463 for Regulation; Table 27) because the rank tests use source-provided MISCS-12 subscale scores, whereas the reliability analysis uses a stricter complete-item record.

I.3 History Length Effects

The history-length diagnostics test whether log history length is associated with score error or score bias after adjustment for six prespecified covariates that are independent of the withheld MISCS

target (Figure 15, Table 59). Longer histories are associated with higher Knowledge signed bias in 17/44 Holm-significant conditions and with lower Regulation absolute score error in 6/44 Holm-significant conditions. History length correlates with score error, and rank ordering on the Regulation subscale remains within the equivalence margin and has no Holm-significant condition.

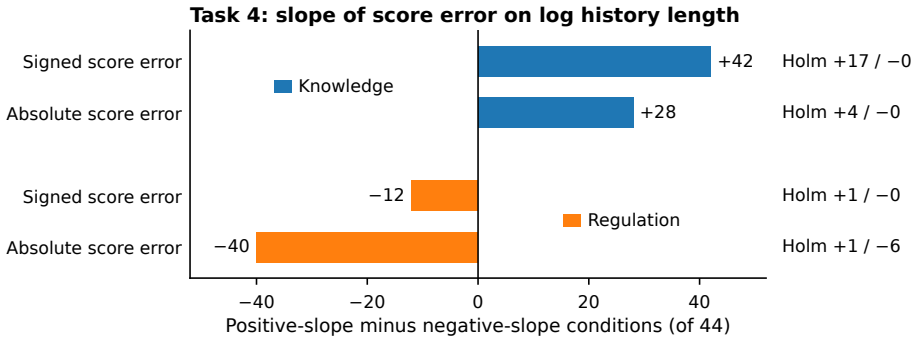

Figure 15: Longer conflict histories go with higher Knowledge score error and lower Regulation absolute score error (Task 4). For each MISCS subscale and outcome, the bar gives the number of the 44 model–language conditions whose slope of the outcome on log history length is positive minus the number whose slope is negative; slopes come from OLS regressions with HC3 standard errors, adjusted for six participant covariates. Signed score error is the predicted minus the self-reported subscale score (1–5 scale); absolute score error is its magnitude. Holm $+a/-b$: a conditions with a Holm-significant positive slope and b with a Holm-significant negative slope (Holm over the 44 conditions of each subscale and outcome). Table 59 gives the counts and median slopes.

Table 59: Associations of log history length with Task 4 score error, adjusted for six participant covariates.

Target	Outcome	Direction conditions	Holm conditions	Median β [IQR]	HC3 CI excludes 0
Knowledge	Absolute score error	positive 36/44; negative 8/44	positive 4/44; negative 0/44	+0.040 [+0.007, +0.075]	15/44
Knowledge	Signed score error	positive 43/44; negative 1/44	positive 17/44; negative 0/44	+0.095 [+0.042, +0.139]	25/44
Regulation	Absolute score error	positive 2/44; negative 42/44	positive 1/44; negative 6/44	−0.076 [−0.097, −0.057]	31/44
Regulation	Signed score error	positive 16/44; negative 28/44	positive 1/44; negative 0/44	−0.024 [−0.053, +0.030]	8/44

Table 60: Task 4 provenance coverage for parsed MISCS outputs. Complete-item-trace rows have finite parsed item-level coverage for the rank summaries; aggregate-score rows have finite subscale scores, each computed from the 12-value list in the response, for all ranked participants and an audited item trace for the listed number of them. Coverage counts conditions out of 44 in the first two rows and, for each aggregate-score condition, participants with a complete item trace out of ranked participants.

Condition	Coverage	Aggregate-score entries
Complete item trace	40/44	0 subscale entries
Aggregate score	4/44	8 subscale entries
Claude Sonnet 4.6 EN	448/467	2 subscale entries
Llama 4 Scout EN	445/467	2 subscale entries
Llama 4 Scout DE	458/467	2 subscale entries
Llama 3.2 3B DE	220/466	2 subscale entries

J SUBGROUP AUDITS

The Results pool across participants and across English and German prompt conditions. This appendix asks where the same persistence pattern stays invariant. The unit of analysis throughout is a subgroup omnibus family, defined as the set of participant-level paired tests within one task, one conflict type, one model, one language, and one stratification axis, each on the P4−P1 metric delta. The subgroup audit uses prespecified partitions, which fix the denominators before testing.

Subgroup criteria. Nine stratification axes are prespecified: eight participant-level axes and one episode-level axis (conflict strength). Three are demographic (sex, current employment status, current university-student status). Three are self-control capacity bands derived from the P3 capacity-for-self-control items (Hoyle & Davisson, 2016) and stratified at the instrument midpoint (> 4 vs. ≤ 4 on the 1–7 scale) per regulatory operation (initiation, continuation, inhibition). Two are MISCS-12 band stratifications (Knowledge band and Regulation band, each split at the instrument midpoint > 3 vs. ≤ 3 on the 1–5 scale). The ninth is conflict strength at exact observed levels 1–7. For every axis, omnibus tests compare subgroup levels on the P4−P1 metric delta across all three conflict types (persistence, initiation, inhibition), all three Tasks 1–3, and 44 model-language conditions, yielding $3 \times 3 \times 44 = 396$ moderation tests per axis. Task 4 subgroup tests (176 per participant-level non-MISCS axis: 2 subscales $\times$ 2 metrics $\times$ 44 conditions) are reported alongside; Task 4 subgroup inference uses the six participant-level non-MISCS axes, whose strata stay separate from the MISCS comparison score. Holm correction is applied within each axis-by-task cell of 132 tests.

J.1 SUBGROUP AUDIT

Table 61: Subgroup omnibus audit summary. Each row aggregates the prespecified subgroup omnibus tests within an axis after Holm correction within each axis-by-task cell of 132 tests.

Subgroup axis	Audit families	Holm-significant	Smallest p
conflict strength	396	6	0.0001
employment status	396	0	0.0006
student status	396	0	0.0055
MISCS Knowledge band	396	12	<.0001
MISCS Regulation band	396	4	<.0001
self-control continuation band	396	3	0.0001
self-control inhibition band	396	18	<.0001
self-control initiation band	396	27	<.0001
sex	396	2	0.0001

Concentration on capacity axes. Subgroup heterogeneity concentrates on self-control capacity axes (Figure 16; Tables 61 and 62). Across the 396 Tasks 1–3 moderation tests per axis (3 conflict types $\times$ 3 tasks $\times$ 44 model-language conditions), the demographic axes yield almost no hetero-

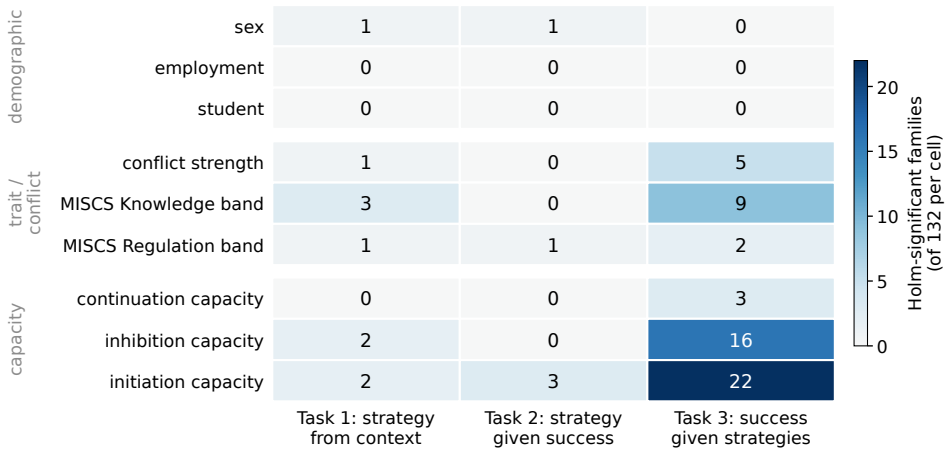

Figure 16: Subgroup moderation concentrates on self-control capacity axes and on Task 3. Each cell counts, out of the 132 subgroup omnibus families for one prespecified axis and task (3 conflict types $\times$ 44 model–language conditions; P4−P1 change in Jaccard for Tasks 1–2 and in absolute success error for Task 3), the families that are Holm-significant within the axis and task and meet the prespecified level-size minima. Rows are grouped into demographic, trait and conflict, and self-control capacity axes; 72 of the 3,564 families meet both criteria.

Table 62: Subgroup omnibus audit by axis and task. Each entry reports Holm-significant subgroup omnibus families within the named axis and task family. Task labels name the prediction target.

Subgroup axis	Task 1 strategy	Task 2 strategy with success	Task 3 success appraisal
conflict strength	1/132	0/132	5/132
employment status	0/132	0/132	0/132
student status	0/132	0/132	0/132
MISCS Knowledge band	3/132	0/132	9/132
MISCS Regulation band	1/132	1/132	2/132
self-control continuation band	0/132	0/132	3/132
self-control inhibition band	2/132	0/132	16/132
self-control initiation band	2/132	3/132	22/132
sex	1/132	1/132	0/132

geneity after Holm correction within each axis-by-task cell of 132 tests (sex 2/396, employment 0/396, student status 0/396), while the self-control capacity bands concentrate the corrected heterogeneity (initiation 27/396, inhibition 18/396, continuation 3/396). Conflict strength (6/396) and the MISCS Knowledge and Regulation bands (12/396 and 4/396) exceed the demographic axes and the continuation-capacity band (3/396) and fall below the initiation- and inhibition-capacity bands.

The cross-task reading supports an interpretation in terms of the process steps. The largest corrected counts occur on Task 3, where success appraisal loads on initiation capacity (22/132), inhibition capacity (16/132), MISCS Knowledge band (9/132), and conflict strength (5/132). Tasks 1–2 show smaller process-aligned moderation: initiation and inhibition capacity account for the largest share of the Holm-significant action-overlap families (7 of 15), and employment and student status stay at 0/132 in each task. Task 4 subgroup tests (1,056 tests across the six participant-level non-MISCS axes; conflict strength is episode-level and does not apply to histories; 176 per axis) yield zero Holm-significant heterogeneity on any axis. The corrected heterogeneity is therefore concentrated where episode-level processes are tested, especially appraisal under different control capacities.

Distribution of capacity moderation. The corrected heterogeneity differs across regulatory operations: it is largest on the initiation-capacity axis (27/396), intermediate on the inhibition-capacity axis (18/396), and smallest on the continuation-capacity axis (3/396). The cross-task pattern is that initiation-capacity differences appear in all three tasks (2/132, 3/132, and 22/132), inhibition-capacity differences are concentrated in success appraisal (16 of 18 families on Task 3), and continuation-capacity differences appear only on Task 3 (3/132).

Because 22 of the 27 Holm-significant initiation-capacity families are Task 3 families, in which the reported strategies are given, this moderation concerns the appraisal step (Section 4.3).

Table 63: Language and subgroup audit families.

Audit family	Unit	Result
Task 1 persona-regime Jaccard language contrast	comparisons	English higher 170/264; German higher 94/264
Task 3 absolute-success-error language contrast	comparisons	English higher 119/264; German higher 145/264
subgroup omnibus families, all prespecified axes	Holm-significant families meeting the level-size minima	72/3,564
self-control initiation band	Holm-significant families meeting the level-size minima	27/396
self-control inhibition band	Holm-significant families meeting the level-size minima	18/396
MISCS Knowledge band	Holm-significant families meeting the level-size minima	12/396
employment status	Holm-significant families meeting the level-size minima	0/396
student status	Holm-significant families meeting the level-size minima	0/396

Prompt-language sensitivity. Prompt language changes the relative metric direction in some comparisons, and the cumulative pattern reported in Section 4 holds in both languages. Task 1 persona-regime Jaccard is higher in English in 170/264 comparisons and higher in German in 94/264; Task 3 absolute success error is higher in English in 119/264 and higher in German in 145/264 (Table 63). English prompts are thus more accurate in most comparisons of both tasks, and the level differences are small (median Jaccard .148 vs. .141). The substantive Task 1 substitution and Task 3 downward appraisal movement both hold in English and German averaged separately. The gap between LLM predictions and reports is shared between English and German in these outputs, and the cross-task asymmetry of Section 4.3 is preserved in either language.

J.2 Within-Family Heterogeneity

Family-median spread is larger on Task 3. Families differ from each other more on the appraisal step than on the action step (Figure 17). Family-median P4−P1 movement on Task 3 absolute

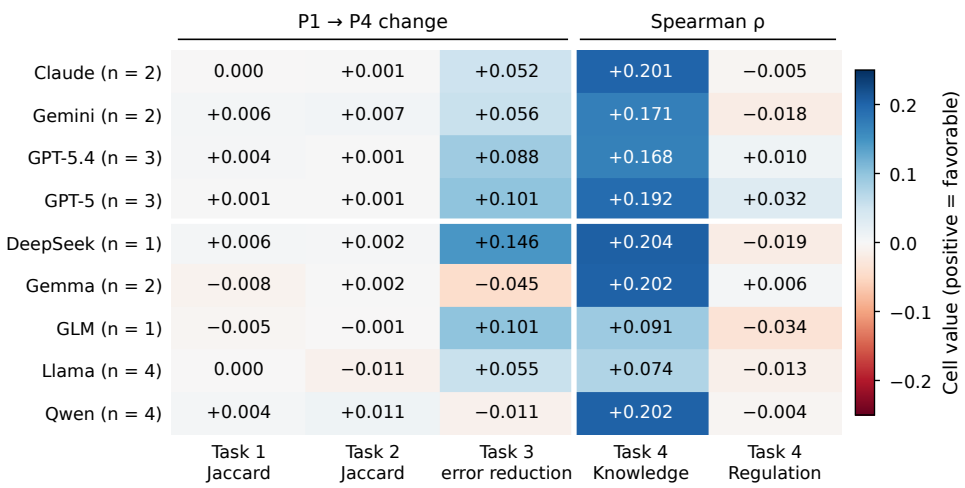

Figure 17: Families differ in Task 3 error reduction and Task 4 Knowledge ranking; differences in Task 1–2 strategy recovery and Regulation ranking are small (persistence). Rows are model families with their number of models n (closed families above the white line); each cell is the median over the family's $2n$ model–language conditions (Table 43 reports the median of per-model means). Tasks 1–2: P4−P1 change in Jaccard; Task 3: P1−P4 reduction in absolute success error; Task 4: Spearman ρ with observed MISCS Knowledge and Regulation scores. Positive values are favorable in every column. GPT-5.4 and GPT-5 are listed as separate families.

success error spans -0.038 (Gemma) to $+0.146$ (DeepSeek), a 0.18-point spread on the success scale; family-median Task 1 Jaccard overlap movement spans only -0.010 (Gemma) to $+0.006$ (DeepSeek), a 0.016-Jaccard spread (Table 43). The family-level signal is concentrated on scalar appraisal movement, while the action-selection step remains close to the same internal/planning-heavy output pattern across families. DeepSeek and GLM lead the favorable Task 3 movement; Gemma is the only family with negative medians on both Task 1 Jaccard overlap and Task 3 absolute success error.

Within-family scale splits. The Task 2 action-step change varies non-monotonically with within-family scale. Qwen 3.5 122B-A10B ($+0.008$ Task 2 Jaccard overlap) and Qwen 3.5 27B ($+0.013$) move favorably, Qwen 3.5 9B (-0.002) moves unfavorably, and Gemma 4 31B ($+0.021$) and Gemma 4 26B-A4B (-0.016) split inside the same release. The pattern recurs on the GPT-5.4 closed series (mini $+0.011$, nano -0.006). Within-family heterogeneity therefore rules out a size-only account of persona-block movement in these outputs.

Region/access partitions. Rows of models developed in China have a higher median Task 2 action movement than the other rows ($+0.007$ vs. $+0.002$), but lower median Task 3 appraisal movement ($+0.018$ vs. $+0.061$); the difference is driven by opposing family rows, with DeepSeek and GLM favorable on appraisal and Qwen weak or negative. Closed-model rows show higher median Task 3 movement than open-weight rows ($+0.079$ vs. $+0.025$), but Task 1–2 movements remain close to zero in both access groups. Each partition jointly indexes family, model release, scale, and serving stack.

Knowledge rank ordering by family. The Knowledge rank signal is also panel-wide. Family-median Knowledge ρ (medians of per-model means) is highest for DeepSeek (0.204), Qwen (0.202), Claude (0.198), GPT-5 (0.192), Gemma (0.188), GPT-5.4 (0.177), and Gemini (0.170); GLM (0.091) and Llama (0.085) trail. No family is below $\rho \approx 0.08$, and no family reaches $\rho = 0.30$; the weak Knowledge association of Section 4.4 recurs at the family level. Three of the four aggregate-score conditions are Llama models (Table 60), and Knowledge is positive in all 40 complete-item-trace conditions.

Table 64: English and German persistence movement agreement across the fixed 22-model panel. Entries summarize the per-model P4−P1 mean for each language and the cross-language correlation. For Task 3, positive values mean lower absolute success error under P4.

Outcome	Models	EN median	DE median	Same-direction models	Cross-language r
Task 1 Jaccard overlap	22	+0.003	−0.000	15/22	0.585
Task 3 abs. error	22	+0.059	+0.051	15/22	0.554

J.3 Language Consistency

Levels are similar across the two prompt languages. In Task 1 persistence at P4, English and German prompts give median Jaccard .148 and .141 and 8.4 and 9.0 strategies per episode, and the same label direction in 84% of the 550 model-by-label pairs; the median Task 4 Knowledge ρ is .175 and .185.

Cross-language directional agreement of persona-block changes is partial. LLM persona-block movement transfers from English to German with moderate agreement, and the agreement is the same on the action step and on the appraisal step. The per-model mean P4−P1 change has the same sign in English and German in 15 of 22 models on Task 1 Jaccard overlap and in 15 of 22 on Task 3 absolute success error (Table 64); cross-language correlations are 0.585 and 0.554. The model-level signal is shared across the two languages tested here, with language-specific variation on each metric; the central pattern (small action movement, larger appraisal movement) holds when each language is averaged separately, and the cross-language correlation does not single out any one task as language-stable.

Pooling across prompt languages. The language results tie back to the central question as follows. The persistence claim is reported as English-and-German pooled because the pattern holds in each language separately and the cross-language correlations are moderate (0.585 and 0.554); the LLMs' failure to reproduce the process is shared across these two languages.

K Statistical Reporting

The main Results report counts, medians, IQRs, and Holm-significant condition counts. This appendix collects the standardized effect sizes and equivalence tests referenced in Sections 4.3 and 4.4.

K.1 Paired Effect Sizes for Persistence

Table 65: Persistence P4−P1 paired Wilcoxon rank-biserial effects for the primary metric of each task. Task labels spell out the prediction direction.

Task	Primary metric	Conditions	Median signed r_{rb} [IQR]	Conditions with $\lvert r_{rb}\rvert > 0.20$
Task 1	Jaccard overlap	44	+0.040 [−0.103, +0.134]	9/44
Task 2	Jaccard overlap	44	+0.043 [−0.092, +0.162]	17/44
Task 3: strategy to success	Absolute success error	44	+0.101 [+0.013, +0.245]	19/44

Paired rank-biserial effects for the primary metric of each persistence task remain small (Table 65). The median signed r_{rb} is $+0.040$ for Task 1 Jaccard overlap, $+0.043$ for Task 2 Jaccard overlap, and $+0.101$ for Task 3 absolute success error under the convention that positive values are favorable; the Task 1–2 IQRs span zero, while the Task 3 IQR $[+0.013, +0.245]$ lies above zero. Conditions

with $|r_{rb}| > 0.20$ number 9/44, 17/44, and 19/44 across Tasks 1–3, with the largest paired effects concentrated on Task 3.

K.2 Bayes-Factor Diagnostics

The TOST and Bayes-factor diagnostics for the Knowledge and Regulation subscales are reported in Table 57. The equivalence margin of ± 0.10 on the Fisher-z scale corresponds to $|\rho| < \tanh(0.10) = .0997$ and is a margin chosen for this study. All Regulation estimates fall inside the ± 0.10 magnitude band; formal TOST equivalence passes in 25 of 44 conditions, and the JZS Bayes factor exceeds 3 in 43 of 44 conditions (median $\mathrm{BF}_{01} = 17.6$). The Knowledge row falls outside the ± 0.10 magnitude band in 37 of 44 conditions, consistent with the corrected positive rank result.

Table 66: TOST equivalence sensitivity at three Fisher-z margins for the Knowledge and Regulation rank associations.

Subscale	Margin	TOST equivalent conditions	$\lvert\rho\rvert < \tanh(\text{margin})$ conditions
Knowledge	± 0.05	0/44	2/44
Knowledge	± 0.10	1/44	7/44
Knowledge	± 0.15	5/44	13/44
Regulation	± 0.05	0/44	39/44
Regulation	± 0.10	25/44	44/44
Regulation	± 0.15	42/44	44/44

The sensitivity rows (Table 66) report TOST equivalence and within-margin counts at three Fisher-z margins, ± 0.05, ± 0.10, and ± 0.15. The within-margin count for Regulation remains high across all three thresholds (44/44 at the wider margins; 39 of 44 conditions satisfy $|\rho| < 0.05$), and formal TOST equivalence rises with the margin (0/44 at ± 0.05, 25/44 at ± 0.10, 42/44 at ± 0.15). At ± 0.05, 42 of 44 Knowledge conditions lie outside the margin, and the within-margin count rises to 13 of 44 at ± 0.15. The headline conclusion uses the ± 0.10 margin, with the stricter ± 0.05 row reported alongside.

K.3 Statistical Reporting Summary

Counts over model-language conditions describe panel recurrence; participant-level paired tests are reported with Holm-significant conditions inside the prespecified family; participant-report distances are computed against the observed participant report; rank tests are reported with both Holm-significant counts and equivalence diagnostics. Inferential claims are made within each conflict type.