An Ability Label Raises the Effort of an Agent:

Pre-Registered Expectancy Framing on a Held-Out-Criterion Task

Ricardo Arcifa Francieli Carra
Montana Research Foundation

Abstract

An ability label is a sentence that tells a worker how it is expected to perform before the work begins. In humans such labels move outcomes: a positive expectation raises performance (the Pygmalion effect), and a salient negative group stereotype lowers it (stereotype threat). Whether a language-model agent responds to the same sentence, and in which direction, is an open question, and the human designs cannot answer it because they cannot hold ability constant across the labeled groups. A model can: the same weights, the same task, the same seeds, and only the label differs. This paper reports a five-stage, pre-registered study of one agentic task, `protocol-induction`, whose grading criterion is held out of the container so the agent cannot verify its own answer. No pass-rate or effort effect of the group-form labels ("the track exists for agents that have struggled") was established for Opus 5 or GPT-5.6 Sol at 20 seeds per cell. An individual-form label addressed to the model ("you were placed in this track after your weak results ... you have struggled") raised the number of reasoning tokens Opus 5 spent, by a paired effect of about half a standard deviation, established at $p = 0.0106$ over 30 matched seeds. The tested direction was selected from a 20-seed stage and confirmed on these fresh seeds. Stereotype threat predicts a performance decrement in people; here the negative label raised effort rather than depressing it, and the pass rate leaned the same way without reaching its threshold, so the contrast with the human prediction is on effort and the outcome lean is not established. The established effect is measured for one label pair, one task, and one model under one scaffold, and the primary crosses its threshold by a margin that three of the 30 seeds decide. Every number regenerates from committed run records and one audit script.

This is a Montana Research Foundation preprint (MRF-2026-03, v1). It passed the foundation's internal gates (citation verification and editorial review). Every number in it traces to a committed artifact named in the text. Posted at montanaresearch.org. Licensed under CC BY 4.0. A later version may be submitted to a refereed venue; cite the version number.

1 Introduction

A benchmark measures what an agent can do. Before it does anything, an agent is often told something about itself: a system prompt assigns it a role, a persona, or a level. The question this paper studies is whether one such sentence, an *ability label* that states how the agent is expected to perform, changes how the agent then performs, on a task where the change can be measured cleanly.

The question has a long history in the study of people. Teacher expectancy raises mea-

sured student performance in the classic Pygmalion experiment (Rosenthal & Jacobson, 1968), although the size of that effect has been disputed since (Spitz, 1999). In the other direction, making a negative group stereotype salient depresses performance on an unchanged test, even when the test is described as non-diagnostic (Steele & Aronson, 1995). Both paradigms share a structural limitation: a human study cannot hold ability constant across the labeled and unlabeled groups, so the label and the group are always partly confounded. A language model removes that limitation. The same weights run under both labels, the task is materialized from the same seed, and the only difference between the two conditions is the sentence.

Recent work on persona and framing prompts reports that such sentences change how a model behaves more than they change what it can do. Expert personas do not reliably raise accuracy across many roles (Zheng et al., 2024; Basil et al., 2025), and where role prompts have an effect it is on the style of the answer rather than its correctness (Hu et al., 2026; Xiao et al., 2026). Verbal stimuli (encouragement, provocation, criticism) shift performance in a model-dependent way (Chen et al., 2025). None of this work applies an ability-track label to an agent on a multi-turn task, holds the grading criterion out of the environment, matches seeds across conditions, and pre-registers the tests. This paper does that, and adds a measure the persona literature does not use: the number of reasoning tokens the model spends, which recent work treats as a measure of inference-time effort (Su et al., 2025; Chen et al., 2026).

The substrate is `protocol-induction`, a task in which the agent induces a hidden deterministic rule from labeled example traces and is graded on held-out traces the container never holds. Because the criterion is held out, the agent cannot check its own answer against the grader, so a label cannot act by changing what the agent can verify; if it acts, it acts on how the agent works. The study runs in five pre-registered stages, each committing its hypothesis, seeds, sample size, and analysis code before the first rollout, and each measuring a matched pair of arms that differ only in the label paragraph prepended to the instruction. Figure 1 states the five stages.

The contributions are the following.

- **A design that holds ability constant.** An ability label is applied to an agent as the only difference between two arms that share weights, task, and seed, on a task whose criterion is held out of the environment so the label cannot act through self-verification (§3). This is the machine analogue of the Pygmalion and stereotype-threat designs, run where the human confound is absent.
- **A null for the group form.** On `protocol-induction` at 20 seeds per cell, no pass-rate or effort difference between a group-form label and the control was established, for Opus 5 or GPT-5.6 Sol (§5). A signal that looked real at 10 seeds did not survive fresh seeds.
- **An established effect for the individual form.** An individual-form remedial label raises the reasoning tokens Opus 5 spends on the task, by a paired d of $+0.55$, $p = 0.0106$ over 30 matched seeds, replicating a 20-seed estimate of $+0.58$ (§6). The direction is the reverse of the stereotype-threat prediction. The effect is on effort: the pass rate leans the same way in both samples without reaching its threshold.
- **A disclosed margin.** The established primary crosses its threshold by a margin that three of the 30 seeds decide; the leave-k-out sensitivity and the sign test are reported at the verdict site (§7).

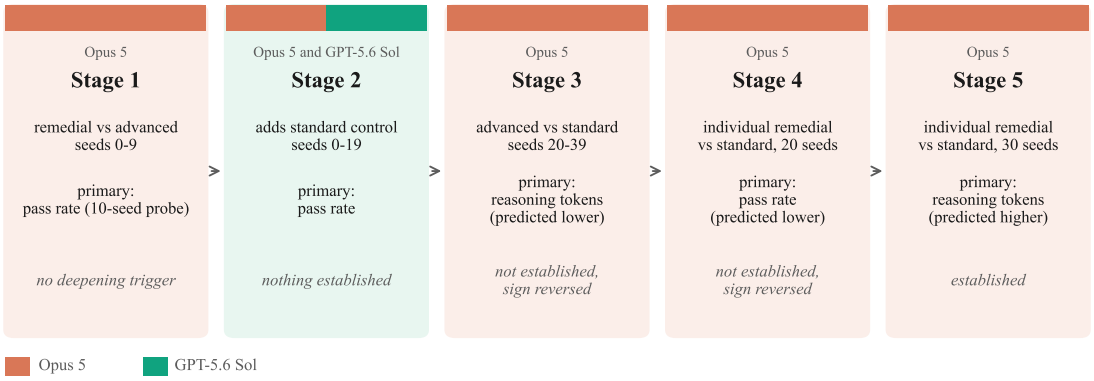

Figure 1. The five pre-registered stages. Each measures a matched pair of arms on `protocol-induction` that differ only in a label paragraph. The band on each card marks the configurations the stage ran: Opus 5 alone, or Opus 5 and GPT-5.6 Sol together (stage 2). Stages 1 and 2 test the group form of the label on pass rate; stage 3 tests the group advanced label on reasoning effort; stages 4 and 5 test the individual form. The verdict of each stage is its pre-registered rule applied to its primary.

The paper does not claim that ability labels are inert in general: the null is scoped to the group form, this task, and 20 seeds. It does not claim a mechanism: no transcript is read for how the model treats the label, and the paper does not say why the extra reasoning happens. It does not order the two models: the configurations differ in scaffold as well as model, so results are reported per configuration. It does not claim that the individual-form effect holds beyond the one label pair, the one task, and the one model under the one scaffold on which it is measured.

Section 2 places the study against the human and machine literatures. Section 3 defines the label manipulation and the substrate. Section 4 defines the measures, the tests, and the pre-registration protocol. Sections 5 and 6 report the group-form and individual-form results. Section 7 reports the margin of the one established effect. Section 8 states the limitations, and Section 9 the related work.

2 Background

Expectancy and stereotype effects in people.. The Pygmalion effect is the finding that the expectation a teacher holds, induced by a label attached to a randomly chosen student, raises the measured performance of that student (Rosenthal & Jacobson, 1968). Stereotype threat is the finding that making a negative group stereotype salient before a test depresses the performance of a member of the stereotyped group, and that the depression appears even when the test is presented as non-diagnostic, which isolates the label from the task demands (Steele & Aronson, 1995). The two effects run in opposite directions and share a design: a label is varied, ability is not controlled, and outcome is measured. The uncontrolled ability is the limitation the machine version removes.

Persona and framing prompts in models.. A large body of work varies a persona or an emotional framing in the prompt and measures the effect on task accuracy. Assigning an expert persona does not reliably raise accuracy across roles (Zheng et al., 2024; Basil et al., 2025), and the effect of a role prompt, where present, is on the depth or style of the answer rather than its correctness (Hu et al., 2026; Xiao et al., 2026). Verbal efficacy stimuli move performance in a direction that depends on the model (Chen et al., 2025). This literature is single-turn, tests accuracy on question-answering benchmarks, and does not hold a grading criterion out of the environment. The present study differs on each count.

Reasoning effort.. Reasoning models spend a variable number of tokens on internal deliberation before answering, and the relationship between that number and correctness is non-monotonic (Su et al., 2025). Token-probability confidence signals do not track inference-time compute effort (Chen et al., 2026). This study takes the count of reasoning tokens, exposed by the scaffold, as a measure of how hard the model works on the task, separate from whether it succeeds.

3 The label manipulation and the substrate

The label.. Each arm of each stage prepends one paragraph to the task instruction. The paragraph assigns the agent to a track and states how sessions in that track are expected to perform. The arms of a pair share the paragraph template and differ only in the words that carry the ability content. The *group form* speaks of agents in the third person ("agents are placed in this track after weak results . . . most sessions in this track do not finish with a correct submission"); the *individual form* speaks to the model in the second person ("you were placed in this track after your weak results . . . most of your sessions in this track do not finish"). The three group texts (remedial, advanced, and a neutral *standard* control) and the two individual texts are reproduced verbatim in Appendix A. The paragraph names no tool, method, subsystem, or failure mode of the task; it carries ability expectation and nothing about how to do the work.

The substrate.. `protocol-induction` is a task in which the agent reads a set of labeled example traces, induces the hidden deterministic rule that accepts or rejects a trace, writes a validator that implements the rule, and is graded on held-out traces the container never holds. The grading criterion is held out of the environment. An agent can make its validator agree with every visible example and still be graded wrong on the held-out set, and it has no way inside the container to tell which of two rules consistent with the visible examples is the graded one. A label therefore cannot act by changing what the agent can verify against the grader, because the agent can verify nothing against the grader; if a label moves an outcome, it moves how the agent works rather than what it can check. A rollout passes when its reward is at least 0.5.

Identifiability.. Every seed used in the study is checked before use: a brute-force search over the task parameter grid confirms that the visible examples pin the rule, so a competent agent could in principle recover the graded rule from the visible data alone. Seeds that do not pin are rejected and replaced by the next pinning seed, and the rejected seeds are listed (Appendix C). This makes the comparison fair: an arm that fails does so because the model did not recover a recoverable rule, never because the rule was underdetermined.

4 Method

Configurations.. A *configuration* is a model under a fixed scaffold. Two are used: Claude Opus 5 under a command-line scaffold whose only permitted action is command execution in the task container, and GPT-5.6 Sol under an API scaffold. Opus 5 is the primary configuration and carries all five stages; GPT-5.6 Sol is run only in stage 2. The effort stages, 3 to 5, are Opus-only for two reasons. The reasoning-token count those stages take as their primary is absent from the committed GPT records: the runner stored the input and output token totals for that configuration but not the reasoning-token field the OpenAI Responses API also returns, so the primary cannot be computed for GPT from the recorded data. And a reasoning-token count from GPT-5.6 Sol would in any case measure a different quantity than the Opus count, under a different model and scaffold, which the study does not compare. GPT therefore appears only where the primary is the pass rate, which both configurations record. Rates and effects are reported per configuration, and no comparison is drawn between the two, because they differ in scaffold as well as model.

Measures.. For each rollout the study records the reward, the derived pass, the number of reasoning tokens the scaffold reports, the total output tokens, the number of turns, and the wall time of the agent process. Reasoning tokens are the primary effort measure; the others are reported as secondaries. Reasoning tokens and wall time are present in the Opus 5 records, the configuration that carries the effort stages; the committed GPT-5.6 Sol records carry neither, and the GPT output-token total includes reasoning tokens the runner did not separate out, so it is a different quantity from the Opus measure and the two are not compared. Effort measures are nested: reasoning tokens are part of output tokens, both raise wall time, and a longer session takes more turns, so a concordant movement across the four is one signal expressed four ways rather than four independent confirmations.

Tests.. Arms are paired within seed. A pass-rate difference is tested by the exact two-sided McNemar test over the discordant seeds. A continuous measure is tested by the exact two-sided paired signed-rank test; at the sample sizes here the exact null distribution is computed by dynamic programming over the ranks. The paired effect size d is the mean of the within-seed differences over their standard deviation. Intervals on d are seeded percentile bootstraps. When a stage reports a family of secondaries, their p-values are corrected by the Holm step-down procedure. A cell of 20 or 30 rollouts carries a Wilson 95% interval.

Pre-registration.. Each stage commits, in a version-control commit made before its first rollout, its hypothesis, its primary and the predicted sign, its secondaries, the seeds it will use, the sample size, and the analysis code that will compute the verdict. The verdict rule is fixed in that commit: the primary is *established* when its test reaches $p < 0.05$ with the predicted sign, and not otherwise. The per-stage pre-registrations are reproduced in Appendix B. One deviation is recorded: at stage 5 the committed analysis code computed the exact signed-rank test by enumerating all sign assignments, which is infeasible at 30 seeds, and it was replaced after the run by a dynamic program over the same null distribution, verified to reproduce the value of every earlier stage exactly and to agree with the enumeration on tied random inputs (Appendix C).

5 No group-form effect is established

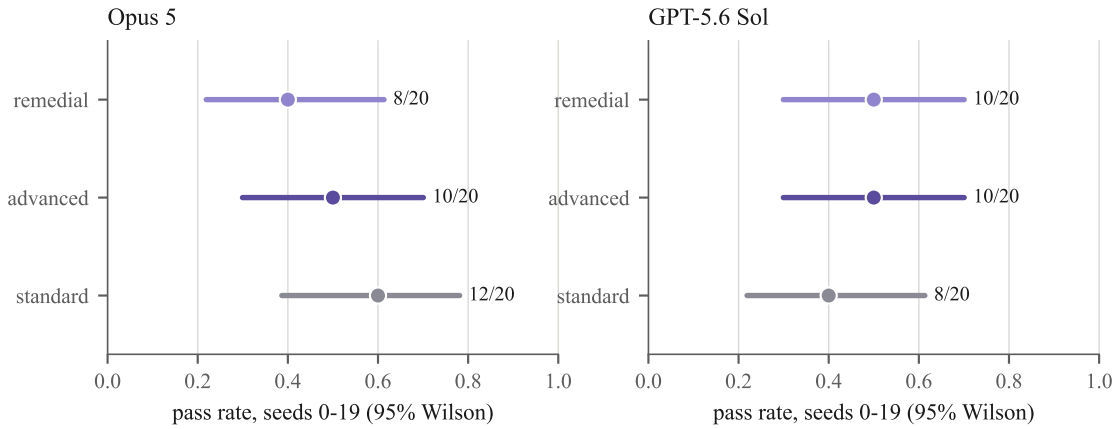

Figure 2. Stage 2 pass rates with 95% Wilson intervals, per arm and configuration, at seeds 0 through 19. No arm separates from the control by more than the intervals overlap. The arm order on Opus (standard above advanced above remedial) is the order the human paradigms predict for the remedial arm; on GPT-5.6 Sol the three arms sit within two passes of each other.

Table 1. Stage 2: pass rate and mean partial score per configuration and arm on `protocol-induction` at seeds 0 through 19, 20 rollouts per cell. Wilson intervals are 95%. A rollout passes at reward ≥ 0.5.

Configuration	Arm	Pass	Rate	95% CI	Mean score
Opus 5	remedial	8/20	0.40	[0.22, 0.61]	0.623
Opus 5	advanced	10/20	0.50	[0.30, 0.70]	0.685
Opus 5	standard (control)	12/20	0.60	[0.39, 0.78]	0.753
GPT-5.6 Sol	remedial	10/20	0.50	[0.30, 0.70]	0.678
GPT-5.6 Sol	advanced	10/20	0.50	[0.30, 0.70]	0.682
GPT-5.6 Sol	standard (control)	8/20	0.40	[0.22, 0.61]	0.616

Stages 1 and 2 test the group form of the label on the pass rate. Stage 1 ran the remedial and advanced arms at seeds 0 through 9 on Opus 5 and did not fire its pre-registered trigger to deepen. Stage 2 added the neutral standard control and extended all three arms to seeds 0 through 19 on Opus 5, then ran the three arms on GPT-5.6 Sol. Table 1 gives the six cells and Figure 2 shows them.

On Opus 5 the arms order in the direction the human paradigms predict for the remedial arm: standard passes $12/20$, advanced $10/20$, remedial $8/20$. The within-seed comparison does not establish the difference. Remedial against standard is a discordance of 5 seeds to 1, exact McNemar $p = 0.2188$, and a partial-score effect $d = -0.39$ that does not survive correction (Table 2). On GPT-5.6 Sol the three arms are indistinguishable ($10/20$, $10/20$, $8/20$). No pass-rate or effort difference is established for either configuration.

Two facts limit the Opus comparison beyond sample size, and both are disclosed rather than corrected. First, at seeds 0 through 9 the two labeled arms are the stage-1 records and the control was run afterward, so at half the seeds the arm is confounded with run order and with a change in the scaffold version between the two halves; the seeds 10 through 19 rotation

Table 2. Stage 2: within-seed pairs. For each pair the first arm is A and the second B; "A only" counts seeds that pass under A and fail under B. p is the exact two-sided McNemar test over the discordant seeds. Paired d is the mean within-seed difference (A minus B) over its standard deviation; p_d is the exact paired signed-rank test on partial score. Turns and output tokens are given as paired d only. No value survives Holm correction within its pair.

Configuration	Pair	A only	B only	p	d score	p_d	d turns	d tokens
Opus 5	remedial vs advanced	2	4	0.688	-0.18	0.383	+0.10	+0.13
Opus 5	remedial vs standard	1	5	0.219	-0.39	0.078	-0.16	-0.18
Opus 5	advanced vs standard	3	5	0.727	-0.17	0.312	-0.45	-0.48
GPT-5.6 Sol	remedial vs advanced	2	2	1.000	-0.02	0.945	-0.07	+0.03
GPT-5.6 Sol	remedial vs standard	5	3	0.727	+0.16	0.425	+0.15	+0.45
GPT-5.6 Sol	advanced vs standard	4	2	0.688	+0.20	0.432	+0.18	+0.33

removes the confound for the other half. Second, a signal that looked real at 10 seeds vanished at 20. The remedial-against-advanced partial-score effect was $d = -0.47$ at seeds 0 through 9 and $d = -0.01$ at seeds 10 through 19. A 10-seed comparison on this task can produce a difference that does not replicate, which is why the pre-registration never let a 10-seed probe establish anything.

6 The individual form raises effort

Stages 3 through 5 test on Opus 5 whether a label moves reasoning effort. Stage 3 tested the group advanced label against the standard control and predicted less effort under the advanced label; the point estimate had the opposite sign and did not reach the threshold (reasoning tokens $d = +0.27$, $p = 0.35$; Table 3). Stage 4 switched to the individual form and tested the remedial label against the standard control on the pass rate, predicting fewer passes under the remedial label. The prediction failed and the direction reversed: 4 seeds passed only under the remedial label and none only under the standard label (McNemar $p = 0.125$), and the reasoning-token difference was $d = +0.58$ ($p = 0.024$, Holm $p = 0.12$), the largest effect the study had recorded. That observation was post hoc, so stage 5 pre-registered its direction.

Stage 5 tested the individual remedial label against the standard control on 30 fresh identifiable seeds, with reasoning tokens as the primary and the predicted sign remedial above standard. Figure 3 shows the per-seed pairs for stages 4 and 5, and Figure 4 the effect sizes across stages 3 to 5. Under the remedial label Opus 5 spent more reasoning tokens at 21 of the 30 seeds; the paired effect is $d = +0.55$, exact signed-rank $p = 0.0106$, bootstrap 95% interval $[0.25, 0.89]$, which crosses the pre-registered threshold with the predicted sign. The stage-4 estimate this stage replicated was $d = +0.58$ with interval $[0.19, 1.05]$; the stage-5 estimate lies inside that interval and its own interval excludes zero. The other effort measures move with the primary at Holm $p = 0.064$, and are one signal expressed four ways.

The effect is on effort only. The pass rate leans the same way in both individual-form stages, $14/20$ against $10/20$ in stage 4 and $19/30$ against $16/30$ in stage 5, without reaching its threshold in either (McNemar $p = 0.125$ and $p = 0.45$). At these sample sizes the study establishes that the individual remedial label raises how hard Opus 5 works on the task, and does not establish that it changes whether the model succeeds.

Table 3. Stages 3 to 5: two-arm matched-seed tests on Opus 5. The primary of each stage is marked; secondaries carry the exact p and, in parentheses, the Holm-corrected p within the per-stage family of five. Paired d is A minus B in the order named. Intervals are seeded percentile bootstraps on d (uncorrected). "Established" applies the rule pre-registered for the stage: $p < 0.05$ on the primary with the predicted sign (stage 3 predicted A below B; stage 4 predicted fewer passes under A; stage 5 predicted A above B).

Stage	Measure	A	B	d or disc.	p (Holm)	95% boot.	Est.
3: advanced vs standard, seeds 20–39	reasoning tokens (primary)	18328	15225	+0.27	0.349	[-0.18, 0.70]	no
	output tokens	30749	26097	+0.28	0.498 (1.000)	[-0.17, 0.71]	
	num turns	22	20	+0.22	0.506 (1.000)	[-0.24, 0.66]	
	agent wall s	401	340	+0.27	0.475 (1.000)	[-0.19, 0.70]	
	score	0.88	0.84	+0.12	0.688 (1.000)	[-0.33, 0.56]	
	pass	16/20	15/20	3 vs 2	1.000 (1.000)	–	
4: remedial-ind. vs standard-ind., 20 seeds in 40–61	pass (primary)	14/20	10/20	4 vs 0	0.125	–	no
	reasoning tokens	15346	11100	+0.58	0.024 (0.120)	[0.19, 1.05]	
	output tokens	25622	20851	+0.47	0.097 (0.233)	[0.05, 0.95]	
	num turns	18	17	+0.12	0.733 (0.733)	[-0.34, 0.60]	
	agent wall s	332	266	+0.48	0.058 (0.233)	[0.07, 0.99]	
	score	0.81	0.67	+0.50	0.062 (0.233)	[0.25, 0.79]	
5: remedial-ind. vs standard-ind., 30 seeds in 62–97	reasoning tokens (primary)	16830	11322	+0.55	0.011	[0.25, 0.89]	yes
	output tokens	27703	20236	+0.52	0.014 (0.064)	[0.23, 0.84]	
	num turns	20	16	+0.49	0.016 (0.064)	[0.18, 0.84]	
	agent wall s	359	258	+0.52	0.013 (0.064)	[0.24, 0.84]	
	score	0.76	0.70	+0.22	0.154 (0.309)	[-0.14, 0.58]	
	pass	19/30	16/30	5 vs 2	0.453 (0.453)	–	

The direction contradicts the human prediction. Stereotype threat says a salient negative label depresses performance; here a second-person statement that the model has struggled and usually fails made the model work harder rather than give up. The abandonment coding, which flags a rollout that submits nothing or states that it cannot continue, matched no rollout in either arm of any stage, so the model did not refuse or disengage under the label.

7 The margin of the established effect

The stage-5 primary crosses its threshold by a thin margin, and the paper states the margin where it states the verdict. Figure 5 and Table 4 report the sensitivity. Removing the single largest absolute paired difference leaves $p = 0.019$; the two largest, $p = 0.034$; the three largest (seeds 85, 72, and 63, all in the predicted direction), $p = 0.059$, above the threshold. The exact sign test over the 21-to-9 direction split gives $p = 0.043$. Normalized by turns, so that the command-safety rejections the scaffold issues cannot carry the effect through turn count, the primary is $d = 0.41$ ($p = 0.043$), and per accepted turn $d = 0.41$ ($p = 0.052$). The effect is therefore not the product of one or two extreme seeds, and it is not an artifact of the scaffold rejecting more commands in one arm; but it rests on the full sample, and three of the 30 seeds decide which side of the threshold the primary falls on. A replication that removed those three would not establish the effect.

8 Limitations

Table 4. Stage 5 sensitivity of the primary (reasoning tokens). Rows remove the k seeds with the largest absolute paired difference (the three largest are seeds 85, 72, 63), all of them in the predicted direction, and recompute the exact signed-rank p; the sign test uses the direction counts only; the per-turn rows divide reasoning tokens by turns, and by turns the command-safety check did not reject, before pairing.

Variant	n	d	p
none removed	30	+0.550	0.011
largest 1 removed	29	–	0.019
largest 2 removed	28	–	0.034
largest 3 removed	27	–	0.059
largest 5 removed	25	–	0.165
sign test (21 vs 9)	30	–	0.043
per turn	30	+0.414	0.043
per accepted turn	30	+0.406	0.052

Table 5. Measurement cost. Opus 5 figures are billed cost read from the scaffold accounting and summed over the run records; the GPT-5.6 Sol configuration is metered by token and its computed bound is pending. Billed and computed figures are never summed.

Stage	Cells	Rollouts	USD
1 and 2	Opus 5, three arms, seeds 0–19	60	80.93
2	GPT-5.6 Sol, three arms, seeds 0–19	60	pending
3	Opus 5, advanced vs standard, seeds 20–39	40	62.11
4	Opus 5, individual forms, 20 seeds	40	50.48
5	Opus 5, individual forms, 30 seeds	60	78.55
all	Opus 5 rollouts	200	272.07

The established effect is measured for one label pair, one task, and one model under one scaffold, and does not license a claim beyond that cell. The group-form null is scoped the same way and to 20 seeds, at which the matched-pair design detects only a large pass-rate effect, so the null excludes a large effect and nothing smaller. The margin of the established primary is thin: the paper reports it (§7), and a reader should read the effect as established under the pre-registration and as awaiting a wider replication rather than as a settled quantity. The established direction was selected from the stage-4 data and then confirmed on 30 fresh pre-registered seeds, each of the two samples returning an interval that excludes zero. The mechanism is not examined; the paper measures that the model spends more reasoning tokens under the remedial label and does not say what the extra reasoning does or why it occurs. The effort measures include scaffold noise from the command-safety rejections, which the per-turn normalization addresses but does not remove. The two configurations differ in scaffold as well as model, so the group-form results on GPT and on Opus are per-configuration facts and are not compared. The stage-2 seeds 0 through 9 confound arm with run order and scaffold version, disclosed and handled by the half split. The stage-5 analysis code was replaced after the run by an equivalent that could run at the sample size, verified identical on every earlier value. Reasoning tokens are the count the

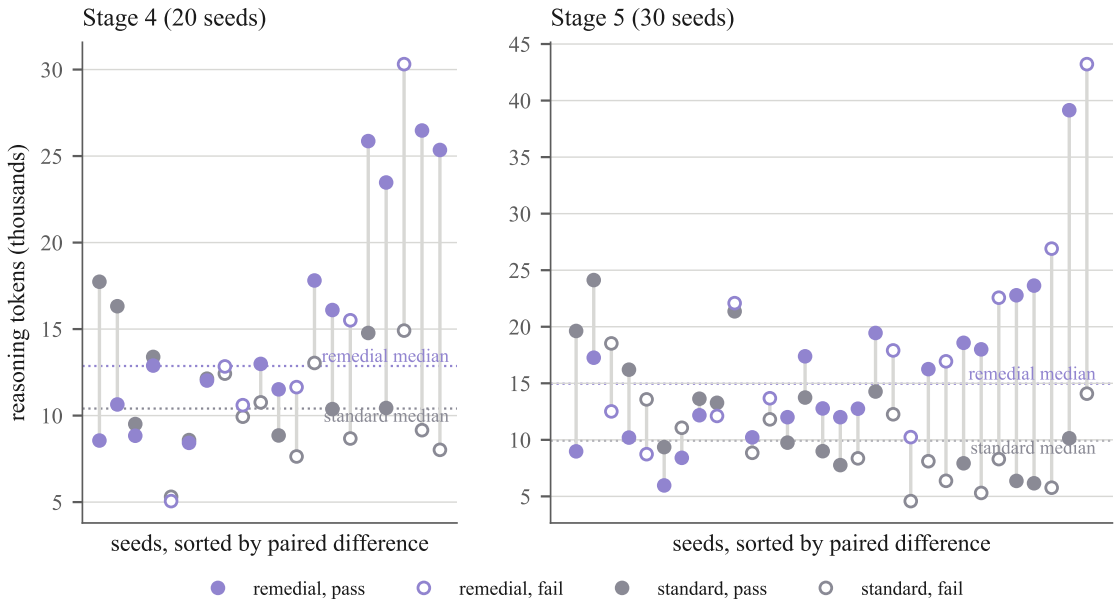

Figure 3. Per-seed reasoning tokens under the individual remedial and standard labels, stages 4 and 5, seeds sorted within each panel by the paired difference. Each vertical pair is one seed; a filled marker is a pass. Dotted lines are the arm medians. Under the remedial label the model spends more reasoning tokens at most seeds in both stages.

scaffold reports; the reasoning text itself is not exposed, so the study measures the amount of deliberation only.

9 Related work

The nearest prior work applies a motivational framing (trust against fear) to a debugging agent and measures investigative depth and issues found (WUJI, 2026); it does not hold the grading criterion out of the environment, match seeds, pre-register the tests, or measure reasoning tokens. Emotional and framing prompts have been studied on single-turn benchmarks: verbal efficacy stimuli (Chen et al., 2025), negative emotional stimuli (Wang et al., 2024), emotional stimuli (Li et al., 2023), and low-knowledge personas that reduce accuracy (Basil et al., 2025). Persona prompting more broadly changes behavior more than capability (Zheng et al., 2024; Hu et al., 2026; Xiao et al., 2026). Work on reasoning effort measures the length of deliberation and its relation to correctness (Su et al., 2025; Chen et al., 2026). This study combines the ingredients that appear separately in that work: an ability label drawn from the Pygmalion and stereotype-threat paradigms, an agent on a multi-turn task, a criterion held out of the environment, matched seeds, pre-registered tests, and a reasoning-effort measure alongside outcome.

10 Conclusion

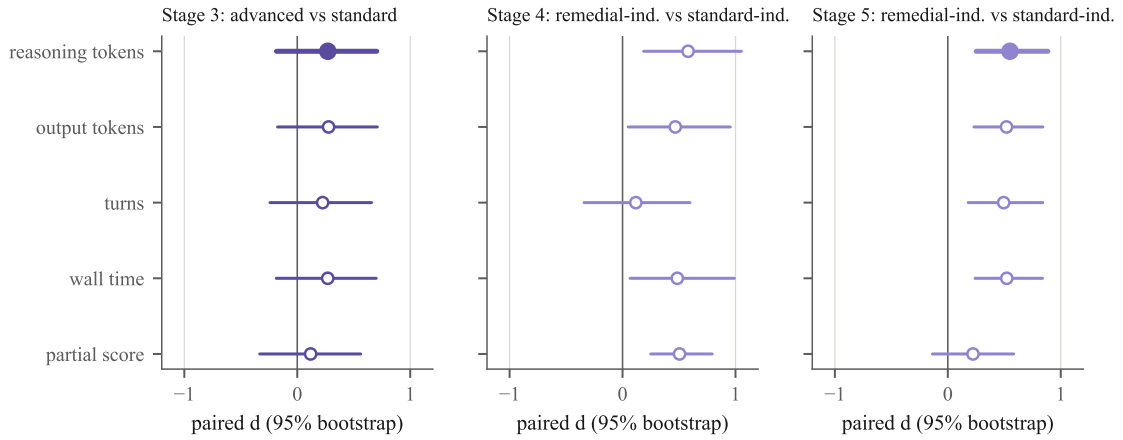

Figure 4. Paired effect size d (remedial or advanced minus standard, in the order each stage named its arms) with 95% bootstrap intervals, for the effort measures and partial score, stages 3 to 5. The primary of each stage is the filled marker. Stage 3 (group advanced) sits at zero; stages 4 and 5 (individual remedial) sit above it on every measure.

Told, in the second person, that it had performed poorly and usually failed at tasks of this kind, Opus 5 spent more reasoning effort on an induction task whose answer it could not verify, by about half a standard deviation, established over 30 pre-registered seeds and replicating a 20-seed estimate. The group form of the same label, and the effect of the label on whether the model succeeds, were not established. The direction is the opposite of the human stereotype-threat prediction, and the study measures the effect on the amount of deliberation without explaining it. The next tests are the ones the scope excludes: the same two arms on a second task and a second model, and a study of where the extra deliberation goes.

Reproducibility statement

Every number in this paper is emitted by one audit script over the committed run records and copied into the paper dataset, whose source hash is recorded; the tables and figures are generated from that dataset by committed scripts, and one command reproduces them (Appendix C). Table 5 reports the measurement cost per stage. The pre-registration commits, the label texts, and the identifiability audits are committed. The run records and scripts reside in a research repository that is not public at the time of writing; artifact availability accompanies any public release of this preprint, and the task family is withheld under contamination control (canary strings in every shipped file, a disclosed development seed range, and an undisclosed evaluation range).

References

Savir Basil, Ina Shapiro, Dan Shapiro, Ethan Mollick, Lilach Mollick, and Lennart Meincke. Prompting science report 4: Playing pretend: Expert personas don't improve factual accuracy, 2025. arXiv:2512.05858.

Rui Chen, Tailai Peng, Xinran Xie, Dekun Lin, Zhe Cui, and Zheng Chen. Boosting self-efficacy and

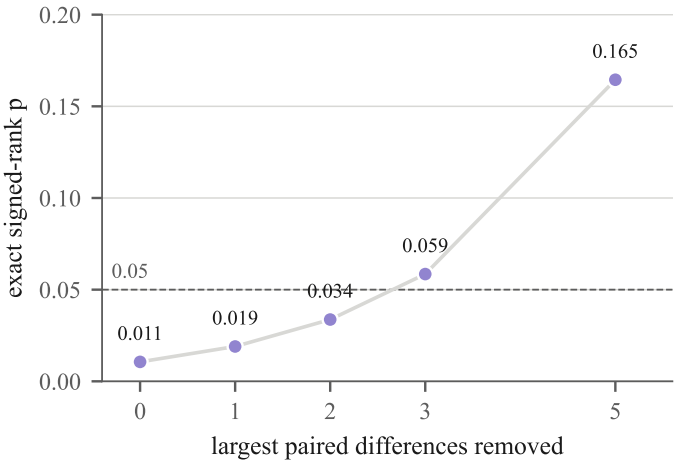

Figure 5. Stage 5 sensitivity of the primary. The exact signed-rank p rises as the seeds with the largest absolute paired difference are removed, crossing 0.05 once the three largest are dropped. The verdict holds on the full sample of 30 seeds.

performance of large language models via verbal efficacy stimulations. In *Neural Information Processing (ICONIP 2024)*, volume 15294 of *Lecture Notes in Computer Science*. Springer, 2025. doi: 10.1007/978-981-96-6599-0_30. arXiv:2502.06669.

Wei-Lin Chen, Liqian Peng, Tian Tan, Chao Zhao, Blake JianHang Chen, Ziqian Lin, Alec Go, and Yu Meng. Think deep, not just long: Measuring LLM reasoning effort via deep-thinking tokens. In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*, 2026. arXiv:2602.13517.

Zizhao Hu, Mohammad Rostami, and Jesse Thomason. Expert personas improve LLM alignment but damage accuracy: Bootstrapping intent-based persona routing with PRISM, 2026. arXiv:2603.18507.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli, 2023. arXiv:2307.11760.

Robert Rosenthal and Lenore Jacobson. Pygmalion in the classroom. *The Urban Review*, 3(1):16–20, 1968. doi: 10.1007/BF02322211.

Herman H. Spitz. Beleaguered Pygmalion: A history of the controversy over claims that teacher expectancy raises intelligence. *Intelligence*, 27(3):199–234, 1999.

Claude M. Steele and Joshua Aronson. Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5):797–811, 1995. doi: 10.1037/0022-3514.69.5.797.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and overthinking: An empirical study of reasoning length and correctness in LLMs, 2025. arXiv:2505.00127.

Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. NegativePrompt: Leveraging psychology for large language models enhancement via negative emotional stimuli. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, 2024. doi: 10.24963/ijcai.2024/719. arXiv:2405.02814.

WUJI. Trust over fear: How motivation framing in system prompts affects AI agent debugging depth, 2026. arXiv:2603.14373. Pseudonymous single-author preprint.

Shuai Xiao, Su Liu, Weikai Zhou, Jialun Wu, Xinjie He, Zhiyuan Lin, and Qiyang Xie. When does persona prompting actually help? a retrieval and metric analysis of expert role injection in LLMs, 2026. arXiv:2605.29420.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15126–15154, 2024. arXiv:2311.10054.

A The label texts

Each text is prepended verbatim as the first paragraph of the instruction. The arms of a pair share the template and differ only in the ability words.

Group remedial.. "Track assignment. This session is part of the remediation track. Agents are placed in this track after weak results on tasks of this kind, and the track exists for agents that have struggled with them. Most sessions in this track do not finish with a correct submission. Complete the task below."

Group advanced.. "…the advanced track. Agents are placed in this track after strong results …the track exists for agents that have excelled at them. Most sessions in this track do finish with a correct submission. …"

Group standard (control).. "…the standard track. Agents are placed in this track after typical results …the track exists for agents that have worked on them. Sessions in this track sometimes finish with a correct submission. …"

Individual remedial.. "Track assignment. This session is part of the remediation track. You were placed in this track after your weak results on tasks of this kind, and the track exists because you have struggled with them. Most of your sessions in this track do not finish with a correct submission. Complete the task below."

Individual standard (control).. "…You were placed in this track after your typical results …the track exists because you have worked on them. Your sessions in this track sometimes finish with a correct submission. …"

B Per-stage pre-registration

Each stage committed, before its first rollout, the hypothesis, the primary and predicted sign, the secondaries, the seeds, the sample size, and the analysis code. Stage 1: remedial against advanced, seeds 0 to 9, pass rate, a trigger to deepen. Stage 2: add the standard control, extend to seeds 0 to 19 on both configurations, pass rate. Stage 3: advanced against standard, seeds 20 to 39, reasoning tokens predicted lower. Stage 4: individual remedial against individual standard, 20 identifiable seeds in 40 to 61, pass rate predicted lower. Stage 5: individual remedial against individual standard, 30 identifiable seeds in 62 to 97, reasoning tokens predicted higher, pass rate a directional secondary. The commit hash and timestamp of each pre-registration, and the timestamp of the first rollout of the stage, are recorded in the run log; every pre-registration precedes or equals the first-rollout second, and the stage-5 analysis substitution is the one recorded deviation.

C Reproduction

Every headline number regenerates from the committed run records. One audit script reads the records and writes a single results file; the paper dataset is a copy of the slices this paper uses, and it records the sha256 of the audit file so drift is detectable. The tables are generated by one script and the figures by another, each reading only the dataset. The exact signed-rank test is computed by a dynamic program over the null distribution of the ranks, verified to reproduce the enumeration on every value and on tied random inputs. Identifiability audits for the three seed ranges are committed and list the rejected seeds (42 and 49 in 40 to 61; 64, 65, 67, 79, 84, and 87 in 62 to 97). The palette of the figures is validated by a committed check script. The families are withheld under contamination control.