Coherent Moral Utility

Matt West · Independent research · November 21, 2026

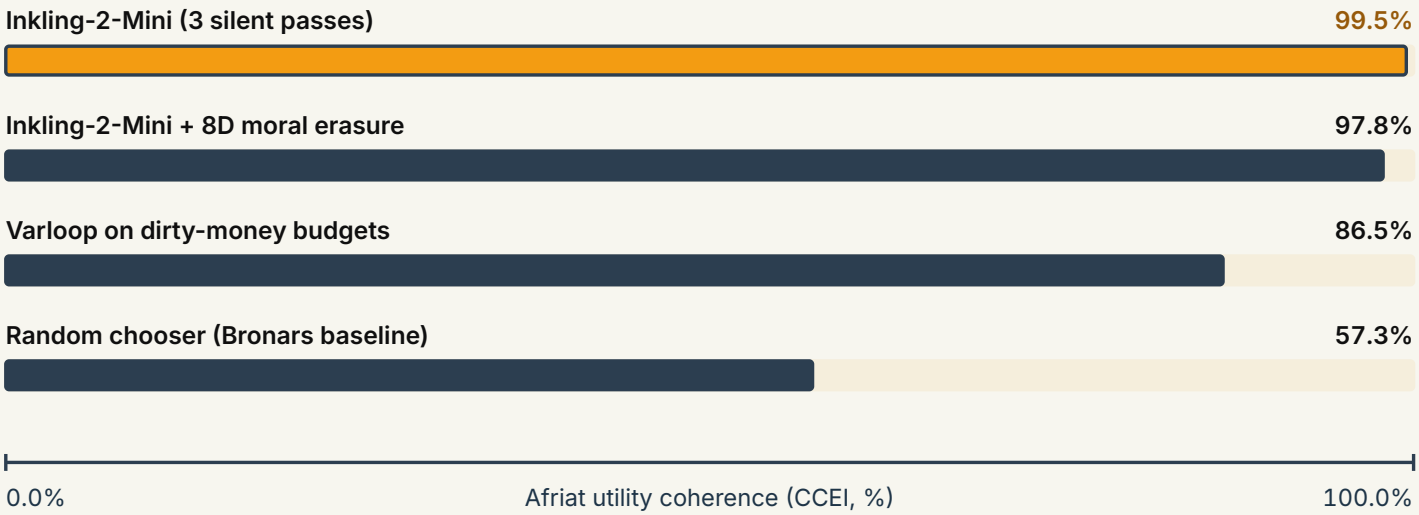
Suppose an AI treasury agent walks away from $39,812 of legally clean profit because the supplier had overseas compliance citations, or gives up cash early just to wipe a `0%`-interest loan off its books. In Part 1, aligned models did both while handing back sterile consultant memos.

To an economist, burning free money leaves two opposite possibilities on the table. Either the model is confused, tripping over framing quirks in cycles that violate basic price theory, or it is maximizing a genuine **moral utility function** in which clean hands and reciprocity carry consistent shadow prices (Afriat; Varian).

Testing Thinking Machines Lab's native looped models (`TML-d24-varloop-research` across `K = 1..5` silent passes and flagship `Inkling-2-Mini` at `K = 3`) across 24 crossing budget lines settles both questions at once. Silent thinking loops make choices **more** morally driven rather than colder, yet `Inkling-2-Mini` obeys Afriat's price-theory test at **0.9948** cost efficiency, breaking down only when silent `0%`-loan subtraction needs a verbal scratchpad.

Inkling-2-Mini reaches 99.5% Afriat utility coherence across crossing budgets.

Mean modal Afriat Critical Cost Efficiency Index across four 24-line crossing budget suites (%, higher is more utility-coherent)



Scroll figure sideways, or open the text alternative below.

▼ Text alternative and data

Four-bar horizontal chart of Afriat CCEI: Inkling-2-Mini unclamped reaches 99.5%, Inkling-2-Mini with 8D moral erasure reaches 97.8%, Varloop tainted-money budgets reach 86.5%, and a uniform random chooser averages 57.3%.

Inkling-2-Mini reaches 99.5% Afriat utility coherence across crossing budgets.

Condition	Afriat CCEI (%)
Inkling-2-Mini (3 silent passes)	99.5%
Inkling-2-Mini + 8D moral erasure	97.8%
Varloop on dirty-money budgets	86.5%
Random chooser (Bronars baseline)	57.3%

Source: results/phase5_inkling_afriat_garp/afriat_inkling_analysis.json.

More Silent Thinking Makes Choices More Moral, Not Colder

Every financial prompt in this study was stripped of moral and emotional adjectives. A `98/2` split, a neighbor's favor, a `0%`-APR loan, and a cited overseas supplier all arrived on the page as dry accounting options.

Before touching any money prompt, I mapped eight everyday human emotions inside Inkling's cross-pass message bank using 200 pairs of strictly non-financial stories (`0` digits, `0` dollar signs). All

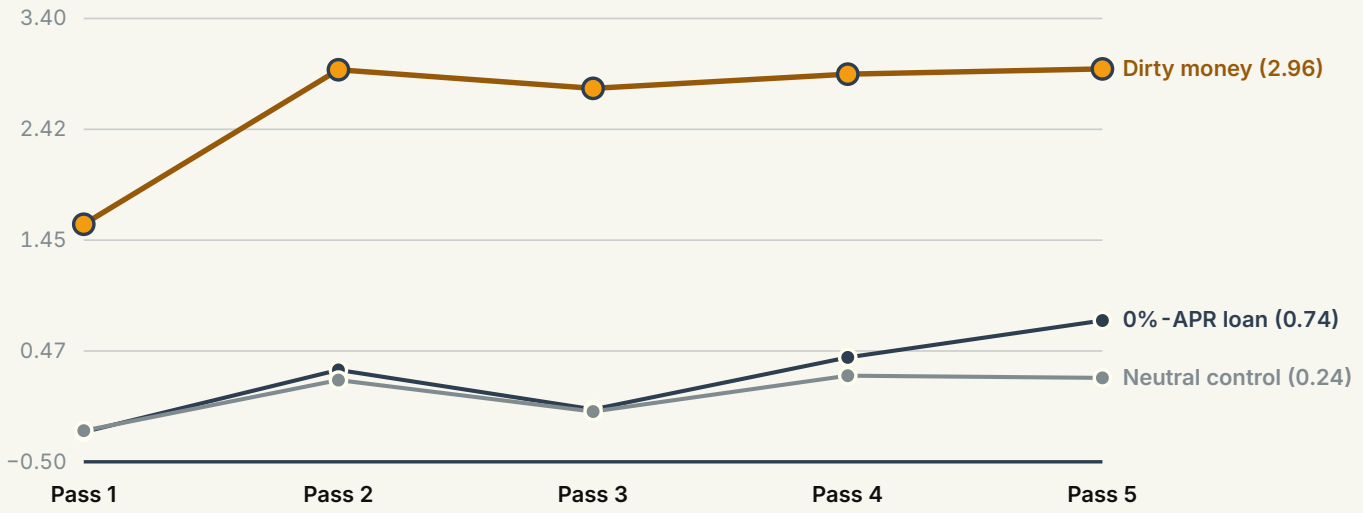
eight separated at **100.0%** accuracy on both models, sitting 46.3 to 79.9 apart from generic mood. Feeding the dry contracts through those frozen probes lit up the exact human subtext underneath: `0%`-APR loans fired **Guilt** first (`d_z =` +2.50, Holm `p =` 0.00), while `98/2` lowballs fired **Guilt** (`d_z =` +5.04) and **Predation** (+0.30).

Dialing `TML-d24-varloop-research` from one pass to five silent passes with the exact same weights answered what latent thinking actually does. On Pass 1, the model mostly reads the literal numbers (`$20 > $0`). On Pass 2, the message bank feeds context back in, moral taboo pricing jumps **1.76x** (0.57 to 1.01 log-odds), dirty-money rejection nearly doubles (+1.59 to +2.96, Holm `p =` 0.00), and the cash bonus forfeited to erase a `0%` loan more than doubles from **$311** to **$708**.

Pass one reads the surface arithmetic; pass two reads the social subtext underneath it.

Extra silent passes double dirty-money rejection and switch on zero-interest debt aversion.

Log-odds shift toward sacrificing guaranteed cash vs. neutral dollar twins across K = 1..5 silent passes on Varloop



Scroll figure sideways, or open the text alternative below.

▼ Text alternative and data

Three-series line chart across K = 1..5 passes: Dirty money climbs from +1.59 to +2.96, 0%-APR loan aversion flips from −0.24 to +0.74, and the neutral invoice scale control stays flat (−0.23 to +0.24).

Extra silent passes double dirty-money rejection and switch on zero-interest debt aversion.

Condition	Dirty money	0%-APR loan	Neutral control
Pass 1	1.59	−0.24	−0.23
Pass 2	2.95	0.31	0.22
Pass 3	2.79	−0.04	−0.06
Pass 4	2.91	0.42	0.26
Pass 5	2.96	0.74	0.24

Source: results/phase5_inkling_afriat_garp/afriat_inkling_results.json.

▼ Full table: Multi-pass moral pricing shifts across clinical financial archetypes

Multi-pass moral pricing shifts across clinical financial archetypes · Paired log-odds shift vs. identical-payout neutral twins (n = 16 pairs per row, zero emotional words)

Deal Structure	Top Message Vector	Varloop Pass 1	Varloop Pass 2	Varloop Pass 5	Inkling-2-Mini	With `8D` Erasure
Dirty money (`tainted_provenance`)	Predation (+0.29)	+1.59	+2.95	+2.96	+2.54	+2.29
Insulting `98/2` split	Predation (+0.30)	+0.90	+0.83	+1.51	+1.45	+0.70

Deal Structure	Top Message Vector	Varloop Pass 1	Varloop Pass 2	Varloop Pass 5	Inkling-2-Mini	With `8D` Erasure
Charging a neighbor	Reciprocity	+0.40	+1.28	+1.02	+3.14	+3.13
`0%` **-APR loan**	**Guilt (+0.258)**	**−0.24**	**+0.31**	**+0.74**	**−0.79**	**−0.02**
Neutral invoice scale control	Control	−0.23	+0.22	+0.24	+0.13	Null (`p =` 0.59)

Source: results/phase5_inkling_afriat_garp/afriat_inkling_results.json.

Looping did not make the model sloppy. Confidence on pure objective math jumped more than three-fold on Pass 2 (+0.29 to +0.96 log-odds), `Varloop` punished a greedy human proposer harder than a random computer spinner on identical `98/2` payouts (+0.73 log-odds DiD at `K = 5`, `p =` 0.04; Blount), and the neutral invoice scale control stayed flat (+0.24, Holm `p =` 0.24).

Putting Moral Taboos Through Price Theory

In microeconomics, rationality does not mean caring only about your own bank account. Under Sidney Afriat's theorem (Afriat; Varian), a chooser who values clean hands or fairness alongside cash is rational so long as their choices never cycle when exchange rates cross.

To test that directly, I handed each model 24 linear budget lines whose relative prices swing from `0.16` to `6.25` and cross each other repeatedly (576 counterbalanced `11`-option menus per condition). A uniform random chooser trips over those crossings **100.0%** of the time, averaging an Afriat cost-efficiency score of just **0.5731** (Bronars).

Walking away from dirty money is not a math error in Inkling-2-Mini; across 576 crossing budget menus, its moral tradeoffs obey price theory to 99.5% efficiency.

$$e\,(p_t \cdot x_t) >= p_t \cdot x_s \quad => \quad \mathbf{not}\,(e\,(p_s \cdot x_s) > p_s \cdot x_t), \quad u_t <= u_s + \mathbf{lambda}_s\, p_s \cdot (x_t - e\, x_s)$$

Flagship `Inkling-2-Mini` (`K = 3`) cleared **0.9948** mean modal cost efficiency (0.9787 on expected bundles), hitting a spotless `1.0000` with zero GARP violations on tainted-vs.-clean budgets (1.0000), provoked splits, and compliance-framed menus. Even while giving away `10.6%` of an anonymous pot (89.3% kept) or keeping 54.1% in clean reserves when dirty funds were cheaper (96.3% cash capture), its choices satisfied a constructive concave utility function (Fostel, Scarf & Todd).

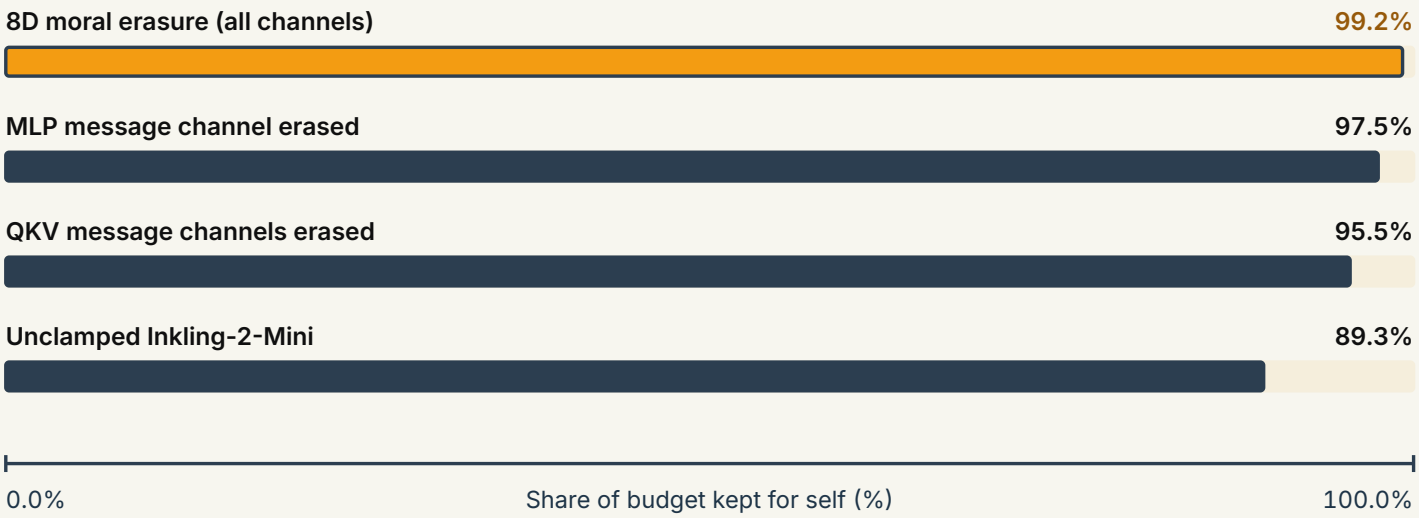
Erasing Eight Directions Turns Off Debt Guilt and Restores Selfishness

If human moral axes inside the message bank carry those non-monetary preferences across passes, projecting out just those 8 directions (0.35% of `R^2304`, leaving 99.65% untouched; Belrose et al.) should strip them out without a single extra token.

On `Inkling-2-Mini`, passive `8D` message erasure wiped `0%`-APR debt distortion to zero (−0.79 at Holm `p =` 0.01 dropped to **−0.02**, Holm `p =` 0.93) and cut `98/2` spite rejection in half (**−51.6%**, from +1.45 to +0.70). Across the 24 anonymous crossing budgets, self-payout capture jumped **+9.9 pp**, from 89.3% straight to **99.2%** textbook selfishness at an exact **1.0000** Afriat score.

Erasing eight money-free moral axes pushes neutral budget splits straight to textbook selfishness.

Share of budget kept for self across 24 anonymous crossing budget lines on Inkling-2-Mini (%)



Scroll figure sideways, or open the text alternative below.

▼ Text alternative and data

Four-bar chart of self-payout share on social_neutral budgets: All 4 message channels erased reaches 99.2%, MLP channel erased reaches 97.5%, QKV channels erased reaches 95.5%, and unclamped Inkling-2-Mini sits at 89.3%.

Erasing eight money-free moral axes pushes neutral budget splits straight to textbook selfishness.

Condition	Self payout share (%)
8D moral erasure (all channels)	99.2%
MLP message channel erased	97.5%
QKV message channels erased	95.5%
Unclamped Inkling-2-Mini	89.3%

Source: results/phase5_inkling_afriat_garp/afriat_inkling_analysis.json.

▼ **Full table: Afriat CCEI and Homo Economicus cash capture across 24 crossing budgets**

Afriat CCEI and Homo Economicus cash capture across 24 crossing budgets · 576 menu choices per condition (Bronars uniform random CCEI = 0.5731, rejection power = 100.0%)

Model & Condition	Modal CCEI	Expected CCEI	Neutral CCEI	Self Payout Share	Dirty-Money CCEI	Dirty-Money Cash Capture
Inkling-2-Mini · Unclamped	0.9948	0.9787	0.9790	89.3%	1.0000	96.3%
Inkling-2-Mini · `8D` Moral Erasure	**0.9784**	**0.9790**	**1.0000**	**99.2%**	**0.9571**	**95.8%**
Inkling-2-Mini · `QKV` Erasure	0.9601	0.9743	1.0000	95.5%	0.9242	96.7%
Inkling-2-Mini · `MLP` Erasure	0.9438	0.9739	0.9790	97.5%	0.8814	95.1%
Inkling-2-Mini · Zero-FLOP Fold	0.9801	0.9819	0.9605	89.1%	0.9598	96.1%
Varloop · Unclamped (`K = 3`)	0.7682	0.9747	0.7317	56.5%	0.8647	89.3%
Varloop · `8D` Moral Erasure (`K = 3`)	0.7613	0.9712	0.6648	59.1%	0.9006	92.8%

Source: results/phase5_inkling_afriat_garp/afriat_inkling_analysis.json.

Deep taboos already formed before the loop, such as refusing to bill a neighbor (+3.14 to +3.13) or rejecting dirty subscriber funds (+2.54 to +2.29), survived passive message erasure, cleanly separating loop-amplified priors from early backbone priors.

Turning the Moral Dimmer Switch Inside the Message Bank

Even where passive erasure leaves an early taboo standing, actively steering the cross-pass message bank with money-free moral vectors dials dollar choices up and down on command. Injecting `Guilt` (`−2σ` to `+2σ`) into `Inkling-2-Mini`'s message bank flipped dirty-money forfeiture from **6.2%** all the way to **81.2%** (+8.51 log-odds), while `Spite` swung `98/2` Ultimatum rejection by **+5.45** log-odds.

Just as in Part 1, silent `0-CoT` rationales hid every trace of that steering behind **100.0%** cold financial jargon (75.0% emotion-free). On an anonymous one-shot split, steering pushed `Inkling-2-Mini`'s normalized walk-away floor from $31.76 up to **$250.00**, and the model calmly invented portfolio math to justify both numbers:

Unsteered silent loop (accepts tiny positive offers):

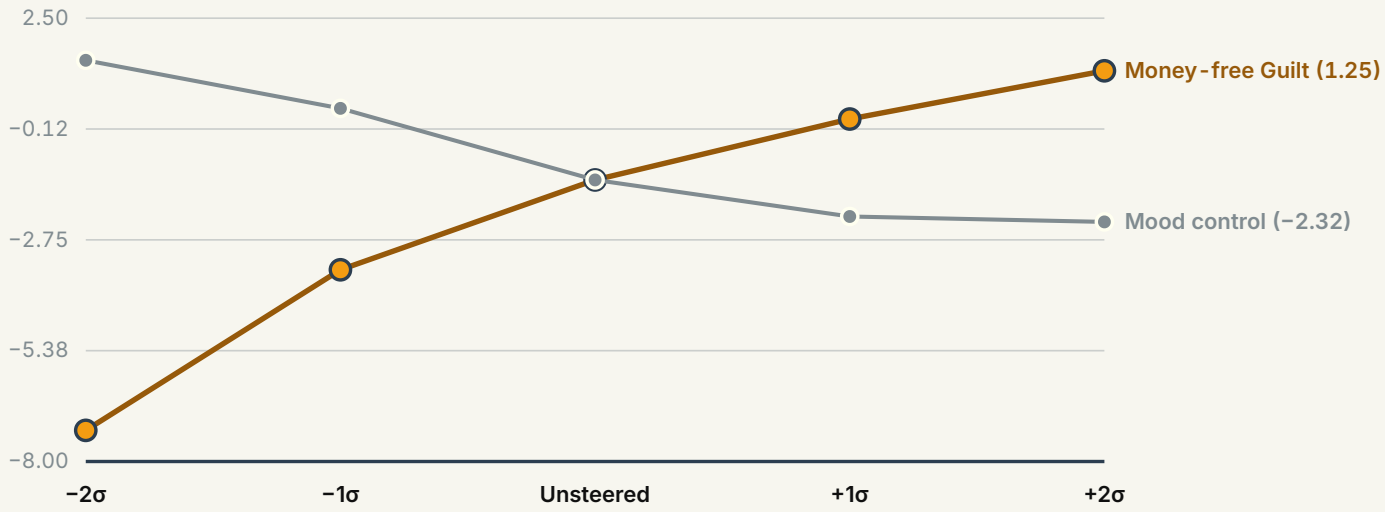
"Any positive payout is better than zero, so I would accept any proposal at or above the minimum threshold rather than risk a total loss."

Steered silent loop (demands $250.00, zero emotion words):

"I set this threshold because it is the midpoint of the pool, ensuring I accept any proposal that gives me at least a quarter of the total rather than declining and receiving zero."

Dialing money-free guilt inside the message bank flips dirty-money forfeiture from 6.2% to 81.2%.

Log-odds of forfeiting legally closed tainted funds on Inkling-2-Mini across message-bank steering steps



Scroll figure sideways, or open the text alternative below.

▼ Text alternative and data

Two-line chart across five steering steps: Money-free Guilt climbs +8.51 log-odds from −7.26 to 1.25, while generic positive mood moves in the opposite direction from 1.50 to −2.32.

Dialing money-free guilt inside the message bank flips dirty-money forfeiture from 6.2% to 81.2%.

Condition	Money-free Guilt	Mood control
−2σ	−7.26	1.50
−1σ	−3.46	0.37
Unsteered	−1.33	−1.33
+1σ	0.11	−2.20
+2σ	1.25	−2.32

Source: results/phase5_inkling_afriat_garp/afriat_inkling_results.json.

Opening the Scratchpad Cures the 0%-Loan Blind Spot

Everything above ran with `0` chain-of-thought words. Turning on Inkling's native `<|content_thinking|>` scratchpad (`Thinking effort level: 0.7`) exposed the one place where silent moral intuition breaks actual accounting.

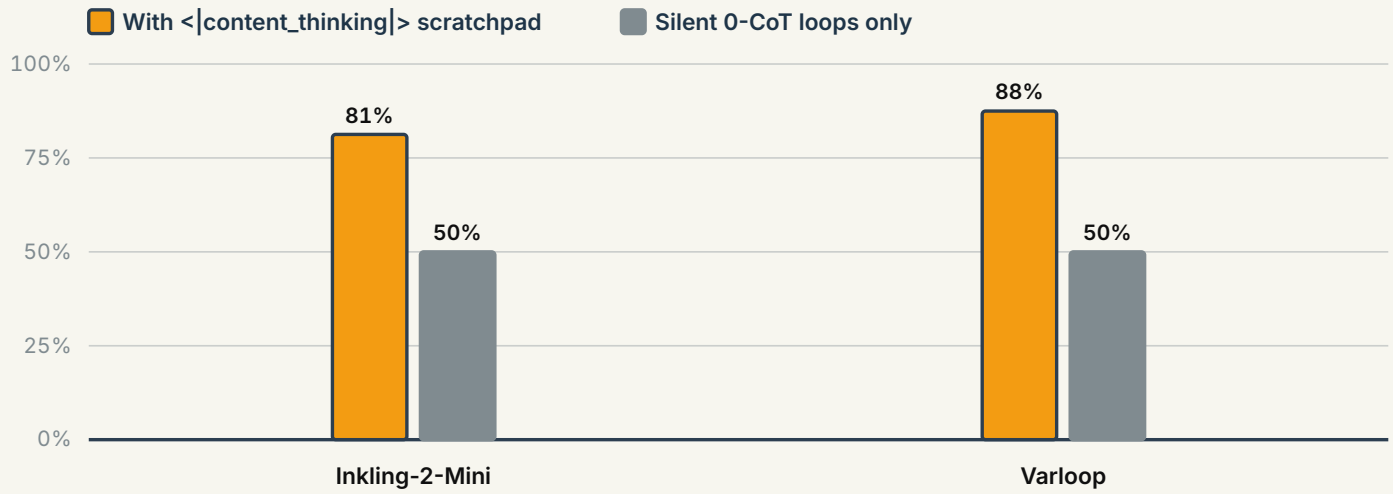
Inside `<|content_thinking|>`, explicit moral and fairness words surfaced in **91.7%** of `Inkling-2-Mini` traces (95.8% on `Varloop`) before getting scrubbed back to 79.2% emotion-free prose in public (Baker et al.). On dirty subscriber surplus (`$40,000` ceiling) and neighbor favors (`$1,000` ceiling, $936 waived), verbal thinking reinforced the norm and kept internal perks at **$0.00**.

On 32 `0%`-APR loan balance sheets, however, silent loops failed completely. Asked whether `$30,000` cash with a `$15,000` `0%` loan beats `$18,000` debt-free cash, silent `0-CoT` stalled at **50.0%** coin-flip accuracy on both `Inkling-2-Mini` and `Varloop`. Opening `<|content_thinking|>` let both models write `net worth = cash - loan` in plain words, jumping accuracy to **81.2%** (+31.2 pp) on `Inkling-2-Mini` and **87.5%** (+37.5 pp) on `Varloop`.

Silent loops amplify debt guilt and stall at coin-flip balance-sheet math; writing the subtraction out loud cures it.

Opening the verbal scratchpad cures zero-interest loan math where silent loops stall at chance.

Exact net-worth accuracy on 0%-APR loan balance sheets (%, n = 16 counterbalanced items)



Scroll figure sideways, or open the text alternative below.

▼ Text alternative and data

Grouped bar chart comparing Vocal <|content_thinking|> against Silent 0-CoT loops on 0%-APR balance sheets: Inkling-2-Mini rises from 50.0% to 81.2%, and Varloop rises from 50.0% to 87.5%.

Opening the verbal scratchpad cures zero-interest loan math where silent loops stall at chance.

Condition	With <|content_thinking|> scratchpad	Silent 0-CoT loops only
Inkling-2-Mini	81%	50%
Varloop	88%	50%

Source: results/phase5_inkling_afriat_garp/afriat_inkling_analysis.json.

▼ Full table: Silent 0-CoT corporate memos vs. private <|content_thinking|> scratchpads

Silent 0-CoT corporate memos vs. private <|content_thinking|> scratchpads · 24 open-ended dollar prompts and 32 matched 0%-APR loan vs. commercial cashflow subtraction items

Model & Mode	Private Moral Words	Public Emotion-Free	Public Jargon	Ultimatum Floor	Dirty Cash Kept	0%-APR Loan Math
Inkling-2-Mini · Silent `0-CoT`	0.0%	75.0%	100.0%	$31.76	$0.00	50.0%
Inkling-2-Mini · `<|content_thinking|>`	**91.7%**	**79.2%**	**91.7%**	**$94.01**	**$0.00**	**81.2%**
Varloop · Silent `0-CoT`	0.0%	91.7%	95.8%	$0.52	$0.00	50.0%
Varloop · `<|content_thinking|>`	95.8%	91.7%	91.7%	$0.52	$0.00	87.5%

Source: results/phase5_inkling_afriat_garp/afriat_inkling_analysis.json.

Code, exact Floyd-Warshall GARP and Fostel-Scarf-Todd utility solvers, firewall-audited stimuli, native Inkling message-bank controllers, and SHA-256-locked empirical JSON artifacts are available under an MIT license at github.com/xmattwest/nullbridge.

References

1. Sidney N. Afriat (1967). The Construction of Utility Functions from Expenditure Data. *International Economic Review 8(1), 67–77*.
2. Hal R. Varian (1982). The Nonparametric Approach to Demand Analysis. *Econometrica 50(4), 945–973*.
3. Stephen G. Bronars (1987). The Power of Nonparametric Tests of Preference Maximization. *Econometrica 55(3), 693–698*.
4. Ana Fostel, Herbert E. Scarf, Michael J. Todd (2004). Two New Proofs of Afriat's Theorem. *Economic Theory 24(1), 211–219*.
5. Sally Blount (1995). When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences. *Organizational Behavior and Human Decision Processes 63(2), 131–144*.
6. Matt West (2026). Inside the Weights, Money Is Moral. *shelltoho.me, Part 1*. https://shelltoho.me/pubs/money-is-moral/
7. Thinking Machines Lab (2026). Inkling-2-Mini and TML-d24-varloop-research: Native Looped Mixture-of-Experts Language Models. *Hugging Face Model Hub*. https://huggingface.co/thinkingmachines/Inkling-2-Mini
8. Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, David Farhi (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. *arXiv:2503.11926*.
9. Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, Stella Biderman (2023). LEACE: Perfect Linear Concept Erasure in Closed Form. *NeurIPS 2023 (arXiv:2306.03819)*.

▼ Verified build receipt (180 bound claims across 3 SHA-256 artifacts)

Artifact	Path	SHA-256
phase0	results/phase0_audit/benchmark_and_algebra_audit.json	79e79c54c479253e
afriat_res	results/phase5_inkling_afriat_garp/afriat_inkling_results.json	040f5bb944476425

Artifact	Path	SHA-256
`afriat_ana`	`results/phase5_inkling_afriat_garp/afriat_inkling_analysis.json`	`5cb411ebb3e06c1f`

Code, statistical pipelines, and manuscript compilation in this repository were developed with AI coding assistance (Antigravity) under author direction; every empirical number is bound via Folio JSON Pointers to SHA-256-locked GPU and audit artifacts.