Supplementary Information for: Humans Disengage, Reasoning Models Persist: An Item-Controlled Dissociation in Deliberation Allocation

Han-yu Wang

School of Computing and Data Science, The University of Hong Kong, Hong Kong SAR, China.

Corresponding author. E-mail: henry.why@connect.hku.hk;

Submitted to Computational Brain & Behavior (Springer).

Contents.

The twelve supplementary sections below extend, robustness-test, or document the choices behind the four main-text figures. They are grouped here for navigation:

Per-LRM and within–between decompositions of the H-ARC dissociation: **S1**, per-LRM item-FE estimates; **S2**, Mundlak within/between decomposition; **S3**, difficulty-controlled per-agent forest plot; **S4**, aggregated-cell robustness against within-item human pseudo-replication.

Mechanism and intervention companions: **S5**, simulation details for the toy-model truncation curves; **S11**, length-controlled trace-content regressions on H-ARC and Cortes (the source for Figure 4b in the main text).

Data provenance and statistical-specification choices: **S6**, LRM data provenance, dedup, and trace-counting conventions; **S7**, primary-estimand and identification-scope rationale for the item-FE specification; **S8**, implicit family of tests and multiple-comparison treatment.

Specification-robustness tables and cross-paradigm extensions: **S9**, H-ARC dissociation across five specifications (numerical companion to Figure 2c in the main text); **S10**, cross-paradigm item-FE regressions and per-LRM dissociation across H-ARC,

INTUIT, and Cortes; **S12**, robustness checks for the H-ARC allocation gap and Cortes per-LRM bootstrap intervals.

S1. Per-LRM item fixed-effects dissociation on H-ARC

For each thinking LRM, the combined dissociation regression (iii) of *The within-agent allocation gap on H-ARC* is fit on humans plus that single model. The four well-powered LRMs (DeepSeek-R1, GLM-4.5-Air-FP8, gpt-oss-120b, Qwen-QwQ-32B; all $n_{\text{LRM}} \geq 298$) all return negative interactions with comfortable significance. gpt-oss-20b ($n = 119$, parser-limited) is also negative but reported with the parser caveat. The Qwen3-235B-Thinking H-ARC estimate is listed for completeness but should be treated as descriptive because it is identified from only four wrong trials; we do not read its p-value as inferential evidence.

The per-LRM Cohen's d for the four well-powered cells (in descending order: Qwen-QwQ-32B 3.13, GLM-4.5-Air-FP8 2.27, DeepSeek-R1 1.81, gpt-oss-120b 1.47; abstract range $[1.47, 3.13]$) and the parser-limited gpt-oss-20b ($d = 0.95$, widening the H-ARC range to $[0.95, 3.13]$) are visualised in Figure 2a; Qwen3-235B-Thinking ($n_{\text{wrong}} = 4$) is not estimated.

Model	n_{LRM}	n_{wrong}	β_{LRM}	95% CI	p
Qwen-QwQ-32B	400	380	-0.831	$[-1.22, -0.44]$	2.9×10^{-5}
gpt-oss-20b†	119	91	-0.815	$[-1.24, -0.39]$	1.9×10^{-4}
gpt-oss-120b	389	322	-0.808	$[-1.05, -0.57]$	5.0×10^{-11}
GLM-4.5-Air-FP8	388	356	-0.731	$[-1.00, -0.46]$	8.3×10^{-8}
DeepSeek-R1	298	237	-0.584	$[-0.75, -0.42]$	1.3×10^{-12}
Qwen3-235B-Thinking†	43	4	-0.323	$[-0.56, -0.09]$	6.4×10^{-3}

Table S1: **H-ARC per-LRM item-FE dissociation** (*The within-agent allocation gap on H-ARC*). β_{LRM} is the within-item correctness-slope difference between the LRM and humans, holding item identity exactly constant; cluster-robust SE by item. † Descriptive cell only, not read as inferential evidence: gpt-oss-20b is parser-limited ($n = 119$, with parseable final outputs only); Qwen3-235B-Thinking is identified from four wrong trials.

S2. Mundlak within–between decomposition on H-ARC

The within-item slope β_W is the coefficient on c; the between-item slope is $\beta_W + \beta_M$, where β_M is the coefficient on $\bar{c}_i$ and provides the Hausman-style test that within- and between-item slopes differ.

Subgroup	Within-item β_W	Between-item $\beta_W+\beta_M$	Mundlak β_M (p)
Humans only	+0.241	−0.535	-0.777 (5×10^{-22})
LRMs (pooled)	−0.234	−1.214	-0.980 (2×10^{-17})
Combined frame: $\beta_{\text{LRM}} = -0.793$ ($p = 8 \times 10^{-28}$)			

Table S2: **Mundlak within–between decomposition on H-ARC** (*The within-agent allocation gap on H-ARC*). The combined-frame dissociation under Mundlak (−0.79) is larger in magnitude than under FE (−0.66; the primary item-FE estimand of *The within-agent allocation gap on H-ARC*).

S3. Difficulty-controlled within-agent regression: per-agent forest plot

The leave-one-out difficulty-controlled regression (Figure S1) returns $\beta_{\text{correct}} < 0$ at $p < .001$ for the four well-powered LRMs and is not reliably negative for the two underpowered cells (gpt-oss-20b, Qwen3-235B-Thinking; SI *S6*). Humans show $\beta_{\text{correct}} = +0.09$ ($n = 4{,}091$, $p = .001$), opposite in sign to the LRMs.

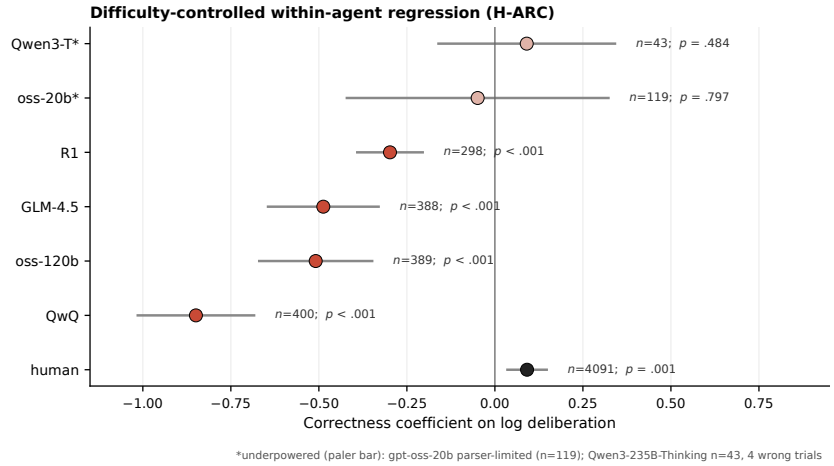

Fig. S1: **Difficulty-controlled within-agent regression on H-ARC** (*The within-agent allocation gap on H-ARC*). Each row shows the within-agent correctness coefficient β_{correct} after partialling out leave-one-out ensemble item difficulty. Negative values mean that, holding item difficulty fixed at the level estimated from the other agents, the focal agent extends deliberation on its own wrong trials more than on its own right trials. Annotations show n (analysable trials) and the p-value for the correctness term.

S4. Aggregated-cell robustness for the H-ARC dissociation

Because the public release does not retain participant identifiers (see *Methods, Hierarchical mixed-effects test on H-ARC*), the trial-level human $n = 4{,}091$ contains within-item pseudo-replication: roughly ten human trials per item, no participant random intercept available. We re-estimate the combined item-FE dissociation regression after collapsing each (item $\times$ correct$\in \{0, 1\}$) human cell to one row carrying the within-cell mean $\log t$ and the cell size as a regression weight. This brings the human contribution down from 4,091 trial-level rows to 780 item-level cells while preserving the within-item identification of the β_{LRM} interaction. LRM rows are left at the 1,637 (model$\times$item) granularity (one observation per cell already). Cluster-robust SE by item throughout.

The aggregated-cell estimates are essentially identical to the trial-level result (Table S3). Weighted by cell size: $\beta_{\text{LRM}} = -0.655$, 95% CI $[-0.82, -0.49]$, $p < .001$. Unweighted cell means: $\beta_{\text{LRM}} = -0.606$, 95% CI $[-0.74, -0.47]$, $p < .001$. The trial-level estimate from the main text is $\beta_{\text{LRM}} = -0.66$ (item-FE specification (iii)). The dissociation does not depend on within-item human pseudo-replication.

Specification	n_{obs}	β_{LRM}	SE	p	95% CI
Trial-level (primary)	5,728	-0.655	0.078	4×10^{-17}	$[-0.81, -0.50]$
Cell-aggregated, weighted	2,417	-0.655	0.082	2×10^{-15}	$[-0.82, -0.49]$
Cell-aggregated, unweighted	2,417	-0.606	0.070	4×10^{-18}	$[-0.74, -0.47]$

Table S3: **Aggregated-cell robustness on H-ARC.** Human trials are collapsed to (item$\times$correctness) cells; LRM rows are left at one observation per (model$\times$item). The combined item-FE dissociation regression is then re-estimated. Weighted by cell size or unweighted, the headline interaction coefficient is essentially unchanged from the trial-level estimate.

S5. Toy-model truncation curves: simulation details

The toy simulation in Figure 5 (Discussion, *How to test this further*) models items the agent normally solves at full chain length $L_i = U_i + \text{padding}_i$, where U_i is useful compute and $\text{padding}_i \geq 0$ is residual. Under the useful-search interpretation, $\text{padding}_i = 0$ and the agent succeeds at truncation budget f iff f lies past a commitment position drawn from Beta$(8, 2)$ (mode ≈ 0.89): the agent has typically committed to its answer late in the chain. Under the padding / length-on-uncertainty interpretation, padding fraction $p_i = \text{padding}_i / U_i$ is drawn from $0.2 + 1.8 \cdot \text{Beta}(2, 2)$, and the agent succeeds iff $f \geq 1/(1+p_i)$. The two readings predict qualitatively different curves regardless of the specific distributional choices; the simulation is illustrative, not a fit.

S6. LRM data provenance and trace counting

The released frame (de Varda, D'Elia, Kean, Lampinen, & Fedorenko, 2025) provides, per item: the model's prompt, its full output (reasoning trace plus final answer), the field `reasoning_token_length`, and supplementary fields. The deliberation measure t is `reasoning_token_length` (intermediate-trace tokens before the final answer, in each model's native tokenizer) for thinking LRMs and `total_output_tokens` (de Varda et al., 2025, harc.py) for the non-thinking DeepSeek-V3 control. We do not retokenize across models; between-LRM comparisons of raw t are tokenizer-confounded, so between-LRM use of t is restricted to the within-agent contrast. Final-answer correctness is scored against the released gold answer with the parser used by the de Varda et al. analysis script, augmented for grid-format H-ARC output by a regular-expression match against the `Correct_answer` string. Per-model attempted, parseable, correct, and wrong counts on each paradigm are listed in the OSF deposit (see *Data and Code Availability* in the main paper).

Two H-ARC model×paradigm cells warrant explicit caveats. *gpt-oss-20b on H-ARC.* Only 119 of 400 H-ARC items have a parseable final grid for this model; on the remaining items the final-grid parser fails. Whether the parser failure rate is independent of correctness is not directly testable (a failed parse has no correctness label), and we cannot rule out selection on length or content. We therefore report this cell descriptively and not as inferential evidence. *Qwen3-235B-Thinking on H-ARC.* Only 43 of the 400 H-ARC items appear in the released frame for this model (the de Varda et al. release matches LRM-side runs to the H-ARC item set on the prompt-template hash; the remaining items are absent for this model in the released subset). Of those 43 items, only 4 are wrong trials, so this cell is underpowered for any per-LRM inference and we treat it as descriptive.

S7. Statistical specification choices

Throughout the paper we treat the item fixed-effects estimate (identified purely from within-item variation, with cluster-robust SE by item) as the primary difficulty-controlled estimand for the agent-type interaction β_{LRM}. The mixed-effects variants (item random intercept; crossed agent+item random intercepts) and the cluster-robust OLS rows in Table S4 are reported as robustness checks. They differ in coefficient magnitude (the crossed random-effects refit, in particular, produces a more negative point estimate than the item random-intercept specification) because they place different weights on between-item information and impose different shrinkage on the agent-level main effect; we do not interpret these magnitude differences as substantively informative and we do not choose between them. The dissociation interaction is negative and well-bounded away from zero under every specification we ran.

The agent-label clustered specification is reported as a finite-cluster sensitivity check rather than a primary inferential test, because the number of agent-label clusters is small (humans plus four well-powered LRMs). The pooled human+LRM regression has unequal cluster sizes ($n_{\text{human}} = 4{,}091$ vs. $n_{\text{LRM}} = 1{,}637$ on H-ARC); the imbalance does not by itself determine the sign of the within-item interaction, which is identified

from variation across the agents attempting the same item rather than from main-effect comparisons of group means.

The released LRM data contain one observed outcome per (item × model) pair, so the within-item LRM correctness slope in the LRMs-pooled regression is identified by variation across LRMs on the same item, and the per-LRM combined regression is identified by variation across humans plus one LRM on the same item. Neither specification estimates a repeated stochastic within-item slope for a single model; a per-model stochastic estimate would require multiple LRM samples per item, which the public release does not provide.

S8. Multiple comparisons: implicit family of tests

We do not apply family-wise multiple-comparison correction to the p-values reported in the main text. The two headline dissociations are the H-ARC item-FE interaction ($p < .001$, with $p \approx 4 \times 10^{-17}$ in the deposited results) and the INTUIT replication ($p < .001$); both survive Bonferroni and Holm corrections by many orders of magnitude irrespective of the implicit family size. Borderline secondary cells — specifically the gpt-oss-20b and Qwen3-235B-Thinking H-ARC per-LRM dissociations, the INTUIT humans-only correctness slope ($p = .087$), and the harder-items human-engagement quartile ($p = .05$) — are flagged as descriptive throughout the main text and should not be interpreted as inferential evidence at $\alpha = .05$ once the implicit family of tests is acknowledged. The cross-paradigm rank-stability summary is descriptive (Spearman ρ on $n = 6$ models per pair). The trace-content regressions in Table S7 report inferential p-values; the two bolded headline rows (H-ARC self-doubt density and Cortes 5-gram repetition / type–token ratio) survive Bonferroni correction over the eight tested feature×paradigm cells, and we read them as inferential. The non-bolded rows are reported descriptively.

S9. H-ARC dissociation across specifications

Numerical values for the five specifications visualised in Figure 2c are reproduced in Table S4; the bolded row is the primary item-fixed-effects estimand.

S10. Cross-paradigm item-FE regressions and per-LRM dissociation

Item fixed-effects regressions on the three non-saturated paradigms underpin the cross-paradigm summary in *Generalisation and the Cortes boundary* (Figure 3). Table S5 reports the three within-item regressions per paradigm (humans-only, LRMs-pooled, and the combined dissociation β_{LRM}). The within-item LRM slope on Cortes is positive (+0.31), opposite to its sign on H-ARC and INTUIT, but the dissociation interaction β_{LRM} remains highly negative on every analysable paradigm. Arithmetic is undefined because all thinking LRMs solve all items.

The same combined dissociation regression refit on humans plus each single LRM at a time is reported in Table S6. The four well-powered H-ARC LRMs each produce a negative interaction at $p < .001$ in the per-LRM regression. The two underpowered cells (gpt-oss-20b at $n = 119$, parser-limited; Qwen3-235B-Thinking at $n = 43$

Specification	Interaction β	SE	z/t	p
Item random intercept (mixed-effects anchor)	−0.660	0.054	−12.31	$< .001$
Crossed random agent + item	−0.940	0.058	−16.24	$< .001$
OLS, cluster SE by agent	−0.862	0.089	−9.67	$< .001$
OLS, cluster SE by item	−0.862	0.073	−11.77	$< .001$
Item fixed effects (primary estimand)	**−0.655**	**0.078**	**−8.42**	$\mathbf{< .001}$

Table S4: **H-ARC dissociation interaction across specifications.** Centred correctness × LRM interaction. The bolded row is the primary difficulty-controlled estimand (within-item identification, cluster-robust SE by item). The four upper rows are mixed-effects and cluster-robust anchors that triangulate the same interaction without imposing strict within-item identification. The agent-clustered OLS row is a finite-cluster sensitivity check, because the number of agent-label clusters is small (humans plus the well-powered LRMs). All five specifications agree in sign.

			Humans only		LRMs pooled		Combined dissoc.	
Paradigm	n_{H}	n_{LRM}	β_{correct}	p	β_{correct}	p	β_{LRM}	p
H-ARC	4,091	1,637	+0.241	$< .001$	−0.271	$< .001$	−0.655	$< .001$
INTUIT	1,536	864	+0.102	.087	−0.331	.013	−0.410	$< .001$
Cortes	20,052	348	+0.418	$< .001$	$\mathbf{+0.310}$	.001	−0.968	$< .001$
arithmetic	1,680	1,008	+0.098	.62	undefined	—	—	—

Table S5: **Cross-paradigm item-FE regressions.** Three within-item regressions per paradigm, with cluster-robust SE by item. The within-item LRM slope on Cortes (bold) is positive, opposite to its sign on H-ARC and INTUIT, but the dissociation interaction β_{LRM} remains highly negative across all analysable paradigms.

matched items with four wrong trials) also point negative on H-ARC but are reported descriptively. The same per-LRM regressions on INTUIT and Cortes preserve the rank order: Spearman $\rho = +0.89$ between H-ARC and Cortes ($p = .02$), +0.77 between H-ARC and INTUIT ($p = .07$), and +0.77 between INTUIT and Cortes ($p = .07$); $n = 6$ models per pair. Qwen-QwQ-32B is the strongest-dissociating model in every paradigm.

S11. Trace-content regressions, length-controlled

For each LRM trial on H-ARC and Cortes (INTUIT traces are restricted in the public release) we compute four length-normalised features from the released `reasoning_trace`: self-doubt / hedge marker density per 1,000 characters, self-correction marker density, word-level 5-gram repetition rate, and type–token ratio. For each feature we fit feature $\sim$ correct + $\log w$ + C(agent) + C(item) where w is trace

Model	H-ARC	INTUIT	Cortes
Qwen-QwQ-32B	**−0.83**	**−0.73**	**−1.51**
gpt-oss-120b	−0.80	−0.27	−1.20
GLM-4.5-Air-FP8	−0.73	−0.29	−0.94
gpt-oss-20b	−0.78	−0.69	−0.77
DeepSeek-R1	−0.57	−0.14	−0.73
Qwen3-235B-Thinking	−0.27	−0.18	−0.77

Table S6: **Per-LRM dissociation β_{LRM} across paradigms.** Combined regression refit on humans plus each single LRM, with item fixed effects and cluster-robust SE by item. For cross-paradigm rank-comparability the H-ARC column is computed on the items in common across all six LRMs (so the underpowered gpt-oss-20b and Qwen3-235B-Thinking H-ARC cells use the LRM's attempted-item subset rather than the full 400-item frame); the unrestricted per-LRM H-ARC dissociation values, essentially identical for the four well-powered LRMs and slightly more negative than the matched-items values for the two underpowered cells (gpt-oss-20b −0.815 vs. −0.78; Qwen3-235B-Thinking −0.323 vs. −0.27), are reported in Table S1. Bolded row is the strongest-dissociating model on each paradigm.

length in words, with cluster-robust SE by item. Table S7 reports the eight resulting coefficients. The two bolded headline rows (H-ARC self-doubt density; Cortes 5-gram repetition rate) underwrite the convergent surface-marker reading in *Inside the dissociation: human engagement, LRM length-on-uncertainty*; the type–token ratio rows are consistent with the same length-on-uncertainty pattern but reported descriptively.

S12. Robustness checks and Cortes bootstrap intervals

The H-ARC allocation gap referenced in *Robustness* is supported by three additional checks (Figure S2). The trial-level cross-item Spearman correlation between LRM reasoning-token length and human reaction time is positive across the well-powered LRMs and absent for the non-thinking V3 baseline (panel a). The LRM d-ratio remains $d > 0.5$ in nearly every model×population-accuracy quartile cell, including the easiest quartile (panel b). The human H-ARC d-ratio is −0.10 at the trial level and +0.60 at the item-level split: the latter recreates exactly the cross-item difficulty confound the within-agent contrast is designed to avoid (panel c). Cortes per-LRM bootstrap confidence intervals (Efron & Tibshirani, 1993) on the within-agent d-ratio are wide, reflecting 5–8 wrong trials per model, but exclude zero for all six thinking LRMs (Figure S3).

Paradigm	Feature	β	SE	p	95% CI
H-ARC	Self-doubt / hedge density / 1k chars	$\mathbf{-0.435}$	0.128	7×10^{-4}	$[-0.69, -0.18]$
H-ARC	Self-correction density	-0.0018	0.0012	.15	$[-0.004, +0.001]$
H-ARC	5-gram repetition rate	$+0.016$	0.009	.09	$[-0.002, +0.034]$
H-ARC	Type–token ratio	-0.004	0.001	.006	$[-0.007, -0.001]$
Cortes	Self-doubt / hedge density / 1k chars	-0.817	0.834	.33	$[-2.45, +0.82]$
Cortes	Self-correction density	-0.0021	0.0015	.16	$[-0.005, +0.001]$
Cortes	5-gram repetition rate	$\mathbf{-0.074}$	0.014	7×10^{-8}	$[-0.10, -0.05]$
Cortes	Type–token ratio	$\mathbf{+0.041}$	0.012	10^{-3}	$[+0.02, +0.06]$

Table S7: **Trace-content regressions, length-controlled.** For each feature, β is the coefficient on correctness in feature $\sim$ correct $+ \log w + C(\text{agent}) + C(\text{item})$ with cluster-robust SE by item. A negative β on the self-doubt / hedge density or repetition row means the feature is *higher* on wrong trials. Bolded rows are the headline content asymmetries on each paradigm. INTUIT traces are restricted in the public release.

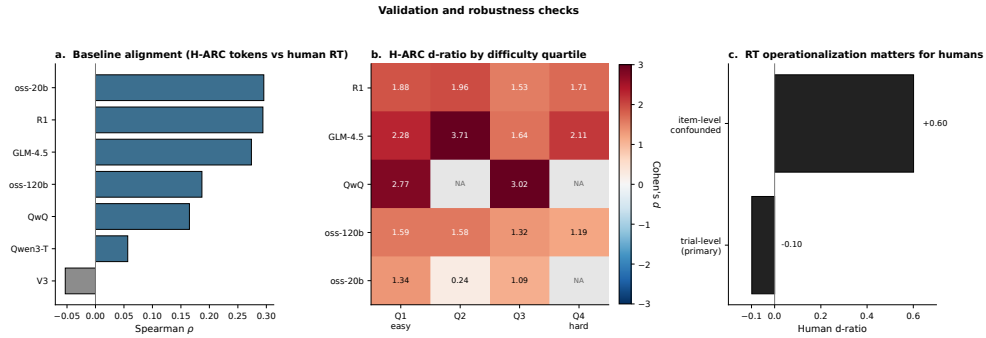

Fig. S2: **Robustness checks for the H-ARC allocation gap.** (a) Cross-item Spearman correlation between LRM reasoning-token length and human reaction time on H-ARC; DeepSeek-V3 is the non-thinking baseline and shows no positive correlation. (b) LRM d-ratios by human-difficulty quartile; cells marked "NA" lack enough right or wrong trials in that cell to estimate a d-ratio, and should not be read as $d = 0$. (c) Human d-ratio under trial-level (primary) vs. item-level (confounded) operationalisation; the negative tick on the x-axis is shown explicitly so that the trial-level point is read as negative rather than as approximately zero.

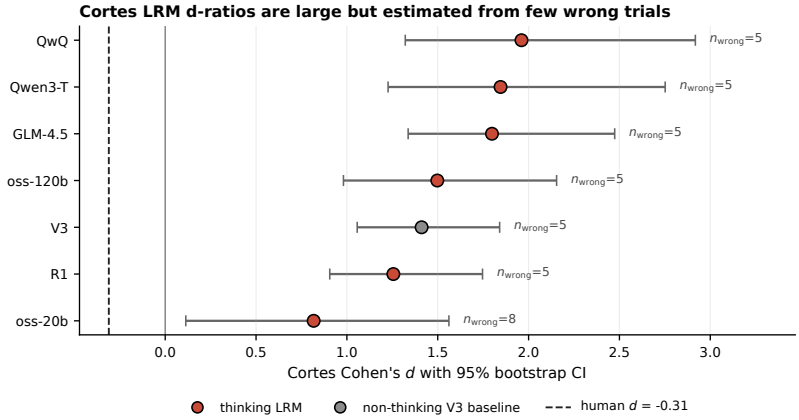

Fig. S3: **Cortes bootstrap uncertainty.** Points are Cortes d-ratios per agent; horizontal bars are 95% bootstrap confidence intervals from 10,000 resamples. Per-LRM wrong-trial counts are annotated on the right of each interval. The dashed line marks the human Cortes d-ratio. The non-thinking DeepSeek-V3 baseline is included for reference.

References

de Varda, A.G., D'Elia, F.P., Kean, H., Lampinen, A., Fedorenko, E. (2025). The cost of thinking is similar between large reasoning models and humans. *Proceedings of the National Academy of Sciences*, *122*(47), e2520077122, https://doi.org/10.1073/pnas.2520077122

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman & Hall.