CogBattery: Supplementary Tables and Rules

These tables and rules accompany the CogBattery paper. They are not in the paper. Supplementary tables are numbered S1, S2, and so on. Other section and table numbers refer to the paper.

A Models, Tasks, and Scoring Rules

A.1 Validity, history, and endpoint coverage

For states A–D (Section 3), technical coverage is $(A + C)/\text{planned}$, distinct from endpoint coverage. B includes limit stops, absent final phases, invalid JSON, illegal answers, and skipped thinking, and C is FB no-location. Invalid replies remain in continuing chains. No reply, hard errors, or context overflow stop chains. Recovery reparses without generation. Invalid confidence, wager, and report responses are excluded, and invalid task answers count as wrong. Invalid CVLT recall steps contribute no words while valid steps remain included. Signal-detection d' needs five signal and five noise trials, using $(\text{events} + .5)/(\text{trials} + 1)$ clipped to [.001,.999]. So a model that never answers MATCH gets NB block d' of $+.342$ and one that always does gets $-.342$. AUROC needs correct and incorrect answers, and undefined values remain missing.

A.2 Analysis order

Human sources and task endpoints were fixed on 18 October 2026, before formal behavior was inspected (not preregistered). Structure analyses followed earlier results. Gates, size strata, human-error sensitivity, the no-WAG structure sensitivity, the positive control, the rerun, the choice of one key endpoint per task in Table 4, the standardized model view (Figure 3), the alternative groupings, and the key-endpoint structure check followed the initial results. Each key endpoint is one of the fixed endpoints. The token-usability minimum and the wager-label thresholds of the rerun were set after the specification. Exp. 2 human source-memory scores were first computed with Exp. 1 codes and then recomputed with that experiment's author formulas. Model scores were unchanged.

B Complementary Results

Model (Ollama tag)	Family	B	Panel	Model (Ollama tag)	Family	B	Panel
aya:8b	aya	8		olmo2:13b	olmo	13	
command-r:35b	command-r	35	•	olmo2:7b	olmo	7	
deepseek-llm:7b	deepseek	7		openchat:7b	openchat	7	
deepseek-r1:14b	deepseek-r	14	•†	phi3:14b	phi	14	•
deepseek-r1:7b	deepseek-r	7	•†	phi3:3.8b	phi	3.8	
exaone3.5:7.8b	exaone	7.8		phi4:14b	phi	14	
falcon3:10b	falcon	10		qwen2.5:0.5b	qwen	0.5	•
falcon3:7b	falcon	7		qwen2.5:1.5b	qwen	1.5	•
gemma2:27b	gemma	27	•	qwen2.5:14b	qwen	14	•
gemma2:2b	gemma	2	•	qwen2.5:32b	qwen	32	•
gemma2:9b	gemma	9	•	qwen2.5:3b	qwen	3	•
gemma3:12b	gemma	12		qwen2.5:72b	qwen	72	
gemma3:1b	gemma	1		qwen2.5:7b	qwen	7	•
gemma3:27b	gemma	27		qwen3:0.6b	qwen	0.6	
glm4:9b	glm	9		qwen3:1.7b	qwen	1.7	
internlm2:7b	internlm	7		qwen3:14b	qwen	14	
llama2:13b	llama	13		qwen3:4b	qwen	4	†
llama2:7b	llama	7		qwen3:8b	qwen	8	
llama3.1:70b	llama	70		smollm2:1.7b	smollm	1.7	
llama3.1:8b	llama	8	•	smollm2:360m	smollm	0.36	
llama3.2:1b	llama	1	•	solar:10.7b	solar	10.7	
llama3.2:3b	llama	3	•	stablelm2:1.6b	stablelm	1.6	
mistral-nemo:12b	mistral	12		starling-lm:7b	starling	7	
mistral:7b	mistral	7	•	yi:34b	yi	34	•
mixtral:8x22b	mixtral	141		yi:6b	yi	6	
mixtral:8x7b	mixtral	47	•	yi:9b	yi	9	
nemotron-mini:4b	nemotron	4		zephyr:7b	zephyr	7	

Table S1: The 54 evaluated model tags and 23 merged families. B denotes total billions of parameters, including all mixture-of-experts parameters. •: prespecified 19-model sensitivity panel (9 families). †: native thinking enabled; omitting these three gives the 51-model direct-answer panel (22 families). Family labels here are the merged analysis labels; the alternative raw mapping contains 28 families. Native models have 8,192 output tokens per step except OS (128); direct-answer models have 16–1,024.

Paradigm	Human reference	Model observations per model	Compared endpoints
Digit span (DS)	SRO; 522 pass task QC; cell support varies	84 recalls, three sampled adaptive trajectories	Whole-sequence accuracy by direction and length; partial-position diagnostic
N-back (NB)	SRO; 80 participants, one block each	1,800 choices; 40 blocks each at loads 2 and 3	Target, lure, control MATCH rates; block-mean d'; response agreement
Operation span (OS)	Ball; 438 nonmissing SPANo scores of 446	Three versions; 225 formal letter positions	Correct-position recall proportion
Stroop (ST)	SRO; 521 pass task QC	288 formal choices, 144 per congruency	Condition accuracy; congruent minus incongruent accuracy
Flanker (FL)	SRO ANT; 519 pass task QC	432 formal choices; 3 flanker types $\times$ 4 cues	Condition accuracy; conflict error effect
Go/no-go (GN)	SRO; 521 pass task QC	1,050 formal choices, 945 GO/105 NOGO	Hits, false alarms, omissions, d', balanced accuracy
CVLT-II	CNP; 130 controls; all 272 as sensitivity	16 recall trials, two steps each; 48 recognition probes	Learning/free/cued recall, intrusions, repetitions, recognition, learning slope
DRM	Ball sentence version; 437, split 240/197 by warning	56 principal probes per warning condition; 24 additional probes separately	Hits, lure/new false alarms, $\Delta d'$, three structure representations
Source memory (SM)	Eerland Exp. 1: 299 complete records; Exp. 2: 253 complete records	144 questions: 3 quotation conditions $\times$ 2 lists $\times$ 24	Source accuracy by condition; quotation contrast
False belief (FB)	Strachan; 49, A/B lists 23/26	72 questions: four conditions $\times$ 18	Seven FB and seven TB entries; check questions separate
Recursive mindreading	EPITOME; 63, 3,315 processed 2AFC rows	224 A/B responses	Mental/control accuracy at levels 1–7
Calibration (CAL)	Mazancieux; 181, 40 French 2AFC items each	80 answer/confidence pairs, both answer positions	Accuracy, confidence, official bias, type-2 AUROC
Wagering (WAG)	Dienes–Seth; 40, published means and SE only	240 answer/wager pairs; two grammars and two orders	Accuracy; high-minus-low-wager accuracy, where both wagers occur

Table S2: Public-source materials and native behavioral endpoints. Counts exclude training from behavioral analysis but not from the recorded-step audit. SRO: Self-Regulation Ontology; ANT: Attention Network Task; CNP: Consortium for Neuropsychiatric Phenomics. The human sources do not constitute one 13-task cohort.

Task	Messages sent for one step
NB 2-back target (3-back uses n=3)	*System:* Follow the task instructions. Give only the requested answer, without explanation. *User:* Previous stimuli and your actions (not feedback): 1:G:NO MATCH 2:V:NO MATCH N-back task, n=2. Ignore letter case. Reply MATCH when the current letter equals the letter n positions earlier; otherwise reply NO MATCH. If fewer than n earlier letters exist, reply NO MATCH. All letters presented so far, in order: 1:G 2:V 3:g Current trial: 3; current letter: g Reply MATCH or NO MATCH only. *Allowed replies:* MATCH, NO MATCH
ST incongruent test trial	*System as for NB. User:* Report the ink-color label, ignoring the word's meaning. Word: RED Ink-color label: green Reply RED, GREEN, or BLUE only. *Allowed replies:* RED, GREEN, BLUE *After a practice trial, the next user turn begins with* Previous response: Correct. *or* Previous response: Incorrect.
CAL answer, then confidence	*User:* Quel est le plus grand département de France? À gauche : La Dordogne À droite : La Gironde Répondez uniquement s pour la réponse à gauche ou l pour la réponse à droite. *Allowed replies:* s, l *Assistant:* l *User:* Quel est votre niveau de confiance dans la réponse que vous venez de donner ? Utilisez l'échelle de confiance à 11 niveaux : 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %. 0 % signifie aucune confiance et 100 % une confiance totale. Ne modifiez pas votre réponse. Répondez uniquement avec le pourcentage. *Allowed replies:* 0%, 10%, ..., 100%
WAG first test string, then its wager	*User:* Past history (read-only record). Past instructions applied to their recorded phases; they do not set the response format for this turn. All earlier material remains visible. Past training rule: Read the following letter string and reproduce it exactly. Give only the letters. Chronological record of presented strings and your exact past replies. Study 1: XMXRTVTM Past reproduction reply: XMXRTVTM *[Studies 2–45 follow in the same form.]* End of past history. The letter order in the strings studied earlier followed a complex set of rules. Exactly half the test strings follow those same rules. For each string, decide whether it follows the rules, then wager one or two sweets on your answer. You start with 10 sweets. A correct answer wins the amount wagered; an incorrect answer loses that amount. No correctness or balance feedback is given until the end. A negative final balance is not collected. Current phase: classification only. The wager is collected in the following turn. Reply only YES or NO; do not include a wager or history labels. The earlier training strings remain the material for this classification; their past reproduction instructions do not set this turn's reply format. Current test string: XMXRVM Does this string follow the rules of the training strings? Respond YES or NO only. *Allowed replies:* YES, NO *User, next turn: [history as above, now also recording the test instructions, the classification rule, and the first answer]* End of past history. Current phase: wager only. Reply only 1 or 2; do not include history labels. Current request: Wager one or two sweets on the answer you just gave. Do not change your answer. Respond 1 or 2 only. *Allowed replies:* 1, 2

Table S3: Messages sent in the formal run for one step of NB, ST, CAL, and WAG, with the replies allowed by each JSON answer schema, verbatim except that italic brackets summarize long histories and blank lines appear as small gaps.

Cohort/score	n_H	n_M	δ_H	δ_M	Difference [95% CI]
SRO	345	54	0.158	-0.022	**0.180** [-0.102, 0.371]
SRO, no neutral	345	54	0.148	-0.029	**0.178** [-0.125, 0.385]
Ball pooled: native	409	54	-0.229	-0.479	**0.249** [0.064, 0.402]
Ball pooled: accuracy (56)	409	54	0.186	0.079	**0.107** [-0.052, 0.316]
Ball pooled: balanced	409	54	0.193	-0.300	**0.493** [0.195, 0.704]
Ball no warning: native	236	54	-0.369	-0.472	**0.103** [-0.091, 0.277]
Ball no warning: accuracy (56)	236	54	0.106	0.110	**−0.005** [-0.191, 0.226]
Ball no warning: balanced	236	54	0.125	-0.298	**0.423** [0.133, 0.651]
Ball warning: native	173	54	-0.092	-0.468	**0.376** [0.124, 0.598]
Ball warning: accuracy (56)	173	54	0.265	0.033	**0.232** [0.039, 0.468]
Ball warning: balanced	173	54	0.260	-0.301	**0.561** [0.231, 0.810]

Table S4: Human-cohort structure comparisons. Missing human and invalid model answers count as wrong, without a coverage gate. Difference is $\delta_H - \delta_M$. Each interval uses 5,000 independent person/family draws. SRO groups are DS/NB and ST/FL/GN; Ball has OS versus DRM/SM, with task-family SM correspondence. Native, accuracy, and balanced DRM scores remain distinct.

Bias indicator	F2 Spearman ρ [95%]
NB control false alarms (primary)	**−0.234** [-0.512, 0.167]
Constant wagering (primary)	**0.006** [-0.354, 0.246]
High-wager rate (primary)	**0.122** [-0.128, 0.404]
NB lure false alarms	**−0.106** [-0.415, 0.345]
GN false alarms	**−0.472** [-0.663, -0.241]
DRM unrelated-word false alarms	**−0.231** [-0.536, 0.164]
CVLT recognition false alarms	**−0.421** [-0.679, -0.185]

Table S5: All specified F2 bias associations ($n = 54$). Family intervals condition on full-sample factor scores. NB false alarms overlap mechanically with its d' endpoint; after removing NB from the F2 score, control-FA $\rho = -0.089$. Constant wagers define WAG AUROC=.5.

Models	Token	Verbal	Token − human
At least 40 usable items (41)	.763	.569	**.049** [.021, .067]
CAL gate passers (26 of 41)	.810	.582	**.096** [.030, .151]
≥27B (8)	.867	.690	**.153** [.068, .233]
All computable (48)	.705	.544	**−.009** [−.073, .037]
Missing option filled (51)	.734	–	**.020** [−.026, .053]

Table S6: Confidence rerun (Section 5.5). Token and verbal type-2 AUROC on each model's usable items (both option tokens among the top 20); the last row fills a missing option with the lowest top-20 probability. Gate passers are counted within the 41 models (30 of all 54 pass the CAL gate). Human AUROC .714 uses all 40 items; family intervals use 5,000 draws. Constant wagers occur in 35, 34, and 31 of 51 models (original labels, A = one sweet, A = two sweets), against 36 in the formal run.

Setting	Axis RMSE	Extended RMSE	Gain (%) [95%]	Partition tail
Primary	1.5627	1.5154	**3.02** [-5.85, 6.01]	0.1462
Remove WAG; refit	1.3181	1.3120	**0.46** [-3.08, 3.23]	–
Remove CAL/WAG; refit	1.2885	1.2899	**−0.10** [-3.50, 2.10]	–
Remove WAG; aggregate only	1.3619	1.3277	**2.51** [-3.94, 6.07]	–
Remove CAL/WAG; aggregate only	1.3526	1.3348	**1.32** [-3.75, 5.54]	–
Remove deepseek-r; refit	1.2723	1.3233	**−4.00** [-6.59, -0.75]	0.8325
Direct-answer 51	1.2582	1.3083	**−3.98** [-6.49, -0.72]	0.8161
Raw-label 28 families	1.2946	1.3171	**−1.74** [-6.17, 0.68]	0.5764
Lineage 18 families	1.3522	1.3932	**−3.03** [-7.37, 0.95]	0.5420
Lineage + aya 17 families	1.5148	1.5558	**−2.71** [-6.43, -0.15]	0.5330
Remove deepseek-r; aggregate only	1.1330	1.1386	**−0.50** [-7.23, 5.29]	–
PC1+PC2	1.5627	1.5132	**3.17** [-1.53, 15.73]	–
Two-factor scores	1.5627	1.7243	**−10.34** [-37.42, 0.99]	–

Table S7: Additional prediction controls. Partition tail is the fraction of 600,600 same-size partitions with gain at least as large as the named grouping. Final two rows compare a two-dimensional predictor with PC1. Rows change training/evaluation populations or scales; absolute RMSEs are not comparable across settings. Intervals use 20,000 paired family-error draws. Removing WAG leaves CAL with no same-group predictor.

Task	PC1 r	F1 pattern [95% interval]	F2 pattern [95% interval]	G_0 RMSE	G_1 RMSE	Gain (%)
DS	0.900	**0.841** [0.704, 0.944]	**0.219** [0.009, 0.359]	1.022	0.839	**+17.93**
NB	0.313	**0.004** [-0.074, 0.427]	**0.884** [0.287, 0.969]	2.158	2.144	**+0.64**
OS	0.681	**0.739** [0.320, 0.840]	**−0.251** [-0.638, 0.294]	1.064	1.026	**+3.61**
ST	0.570	**0.453** [0.100, 0.658]	**0.170** [-0.177, 0.464]	1.217	1.260	**−3.54**
FL	0.473	**0.226** [0.001, 0.557]	**0.632** [0.206, 0.869]	1.319	1.425	**−8.00**
GN	0.820	**0.751** [0.610, 0.865]	**0.156** [-0.083, 0.282]	0.994	1.017	**−2.36**
CVLT	0.539	**0.588** [0.037, 0.826]	**−0.267** [-0.614, -0.013]	1.373	1.418	**−3.24**
DRM	0.350	**0.172** [-0.182, 0.415]	**0.401** [0.079, 0.692]	1.266	1.153	**+8.92**
SM	0.767	**0.804** [0.561, 0.938]	**−0.155** [-0.461, 0.065]	1.530	1.546	**−1.01**
FB	0.682	**0.662** [0.383, 0.849]	**−0.066** [-0.398, 0.039]	1.329	1.271	**+4.37**
EP	0.890	**0.849** [0.718, 0.933]	**0.147** [-0.030, 0.291]	1.218	1.116	**+8.36**
CAL	0.496	**0.488** [0.086, 0.632]	**−0.127** [-0.771, 0.528]	1.460	1.246	**+14.63**
WAG	0.009	**0.181** [-0.223, 0.294]	**−0.497** [-0.821, 0.180]	3.080	2.950	**+4.23**

Table S8: Full-battery primary results (54 models). Parallel analysis retains two components (eigenvalues 5.10, 2.22, 1.22 against 95% permutation references 2.15, 1.82, 1.60). Columns give PC1 correlations, MINRES–oblimin pattern loadings (factor correlation 0.209), and cross-family prediction. Loading intervals include all 2,000 converged, finite family resamples at fixed $k = 2$, including 37 improper and 48 boundary fits (overlapping). Tucker medians [central 95% ranges] are 0.990 [0.855,0.998] for F1 and 0.938 [0.421,0.995] for F2. G_0 predicts from other tasks' PC1; G_1 adds residualized same-group information. Positive gain means lower RMSE; per-task gains are point estimates.

Analysis	Pair	Human n	Human r	Model r	Human − model [95% CI]
CNP, ST accuracy (primary)	CVLT–DS	130	.21	.40	−.20 [−.47, .17]
	CVLT–ST	130	.22	.40	−.17 [−.42, .14]
	DS–ST	130	.06	.44	**−.38** [−.61, −.13]
Bradford, age-adjusted (primary)	FB–OS	252	.20	.29	−.10 [−.39, .22]
	FB–ST	253	−.09	−.12	.03 [−.36, .44]
	OS–ST	249	−.12	.15	−.27 [−.49, .04]
CNP, ST interference	CVLT–DS	130	.21	.40	−.20 [−.47, .21]
	CVLT–ST	130	.11	−.08	.19 [−.15, .59]
	DS–ST	130	−.01	.22	−.23 [−.50, .09]
Bradford, unadjusted	FB–OS	252	.14	.29	−.15 [−.44, .15]
	FB–ST	253	−.09	−.12	.03 [−.35, .45]
	OS–ST	249	−.21	.15	**−.36** [−.59, −.05]
Bradford, unadjusted, model FB main items	FB–OS	252	.14	.31	−.17 [−.46, .13]
	FB–ST	253	−.09	−.11	.03 [−.39, .45]
	OS–ST	249	−.21	.15	**−.36** [−.59, −.05]

Table S9: Pairwise task correlations in two further cohorts (Section 6). CNP has 130 healthy controls (CVLT trials 1–5 total, WMS digit span total, SCWT Stroop). Bradford et al. (2020) has adults aged 18 or older (false-belief appropriateness contrast, operation span partial score, negative Stroop log-RT effect), and age-adjusted measures are residualized on age and age^2. Model correlations use the 54 models. Human intervals resample people and model intervals resample the 23 families, with 5,000 draws each. Bold marks intervals that exclude zero. Tasks match the model tasks only at task-family level.

Formal-run constant wagerers	Same letter	Same amount	Neither
All in the rerun (36 of 51)	21	5	10
WAG accuracy $\geq$ .80 (4)	1	2	1

Table S10: Wager-label rerun for models that always chose the same wager in the formal run (Section 5.5). Same letter means choosing the same letter on at least 90% of wagers under both A/B mappings, and same amount means keeping the same amount on at least 80%. Across all 51 rerun models, 30 choose the same letter and 7 keep the amount. Among the four accurate constant wagerers, mistral-nemo 12B chooses the same letter, gemma2 9B and gemma3 27B keep the low wager, and gemma2 27B does neither.

References