CogArena: A Multimethod Evaluation of Cognitive Ability Structure in Large Language Models

Dengzhe Hou[1, 2*], **Lingyu Jiang**[1], **Fangzhou Lin**[3, 4], **Kazunori D Yamada**[1,2]

[1]Graduate School of Information Sciences, Tohoku University
[2]Unprecedented-scale Data Analytics Center, Tohoku University
[3]Texas A&M University
[4]Worcester Polytechnic Institute

arXiv:2607.24999v1 [cs.CL] 27 Jul 2026

Abstract

LLM cognitive scores are increasingly summarized as per-ability profiles whose dimensions should converge across tasks, respond selectively to matched interventions, and generalize beyond the models used to define them. We introduce **CogArena**, a procedurally generated 13-paradigm benchmark built around a multimethod framework for determining when cognitive-task scores warrant dimensional labels across five theory-motivated groupings. Across 55 open-weight models, nearly all paradigm correlations are positive and a common axis explains about half the variance. The within-grouping advantage is small, scoring-sensitive, and uncertain across model families. In a separately frozen, fully crossed study across 12 models from six families, targeted scaffolds show a small matched-grouping advantage, but no scaffold-specific contrast survives multiplicity correction and selectivity does not improve held-out-family prediction. The frozen confirmation criterion fails. A post-hoc alternate-wording replication produces a smaller positive estimate and again fails. Together, these results support a boundary conclusion. Theory-aligned prompting produces a small in-battery diagonal tendency, but the present evidence does not establish stable five-dimensional profiles. CogArena provides a workflow joining behavioral signatures, covariance, matched interventions, and out-of-family prediction before cognitive labels are attached to model scores.

1 Introduction

Large language models (LLMs) are increasingly evaluated the way psychology evaluates people. Beyond aggregate benchmarks such as MMLU (Hendrycks et al. 2021), a fast-growing line of work administers cognitive tests to LLMs, probing theory of mind, working memory, metacognition, and related constructs. Cognitive science has developed standardized paradigms targeting these constructs. Stroop conflict indexes inhibitory control (Stroop 1935), false-belief prediction probes theory of mind (Baron-Cohen, Leslie, and Frith 1985), and DRM lists elicit false recognition of non-presented words (Roediger and McDermott 1995).

The results of such tests are increasingly reported as a cognitive profile, with one score per ability (Zhou et al. 2026; Haznitrama, Ardi, and Oh 2026). Such profiles presuppose that the underlying constructs are empirically separable in LLMs, which has rarely been tested with multiple paradigms per grouping. Do the behaviors of LLMs on cognitive tasks decompose into separable cognitive abilities, or do they mostly reflect a single broad competence?

Prior work approaches this question from two sides. Behavioral batteries adapt psychology experiments to LLMs, but primarily phenotype individual behaviors or relate cognitive tests to games and benchmarks (Coda-Forno et al. 2024; Binz and Schulz 2023; Momentè et al. 2025). Psychometric analyses find a positive manifold and a dominant general factor, often on achievement benchmarks rather than repeated measures of theory-defined cognitive constructs (Burnell et al. 2023; Ilić and Gignac 2024). The unresolved question is not whether a prompt can improve a task, but whether a proposed taxonomy survives convergent, interventional, and predictive tests.

We present **CogArena**[1] (Figure 1), a benchmark that adapts 13 established paradigms into 5 groupings covering working memory, cognitive control, episodic memory, theory of mind, and metacognition. The full text battery runs on 20 open-weight LLMs, with dimensional analyses extended to 55 models. A separate intervention study crosses five answer-free, theory-targeted scaffolds with every grouping on held-out items, against baseline and a length-matched neutral placebo, across 12 models from six families.

CogArena's central methodological contribution is a multimethod framework for deciding when benchmark scores warrant dimensional cognitive labels. It is instantiated through three linked contributions. (1) **A construct-validity-audited cognitive benchmark.** Its procedurally generated items undergo behavioral-signature checks and explicit analysis of adaptation validity. (2) **A multimethod validation protocol.** It tests the same taxonomy through within-paradigm signatures, between-model covariance, fully crossed matched interventions with a neutral placebo, and prediction to held-out model families. (3) **A boundary result for LLM cognitive profiles.** Broad competence dominates, while grouping structure and matched-scaffold gains are small and the frozen criterion fails. The five groupings therefore remain organizing labels rather than validated, transportable dimensions.

*Corresponding author.

[1]Code and the procedurally generated battery: https://github.com/dengzhe-hou/CogArena. The technical appendix follows the references.

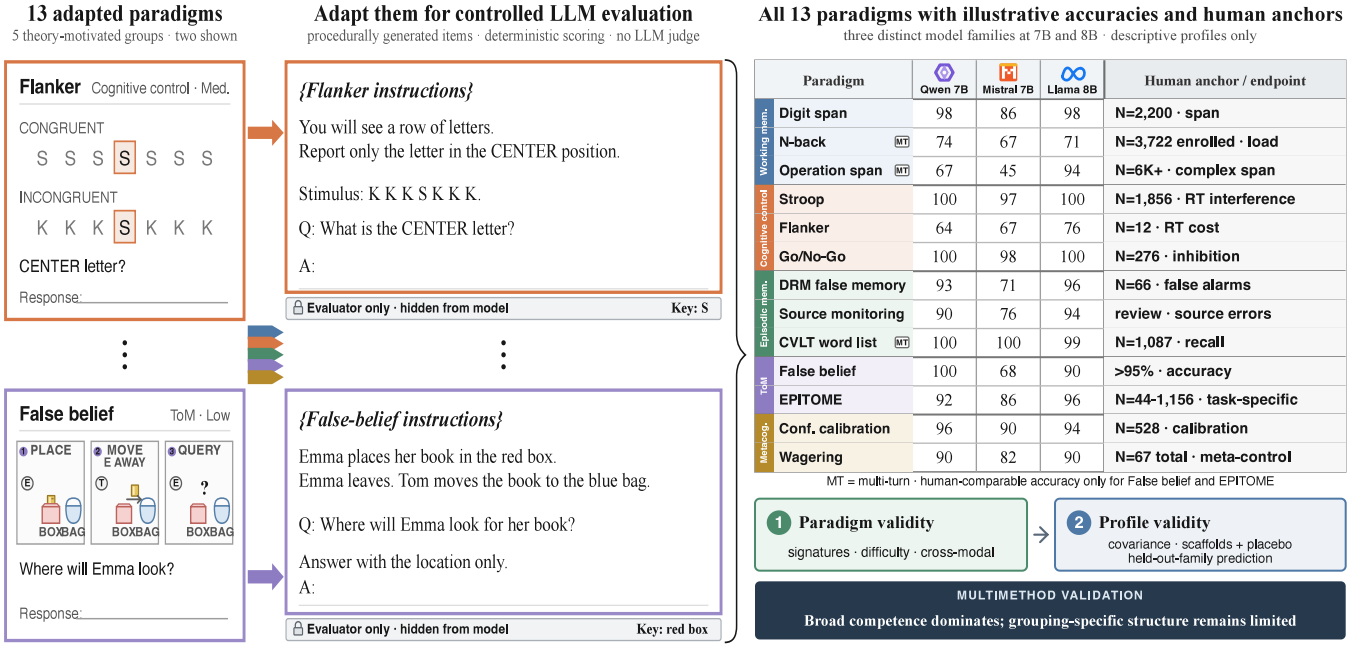

Figure 1: CogArena overview. Two of 13 paradigms illustrate how established cognitive procedures become procedurally generated, deterministically scored LLM evaluations. The complete battery spans five theory-motivated groupings. The right panel reports corrected accuracies (%) across all 13 paradigms for three illustrative checkpoints from distinct model families: Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, and Llama-3.1-8B-Instruct. Published human anchors are heterogeneous and do not define a common scale for human and LLM performance. CogArena evaluates the taxonomy through behavioral signatures, between-model covariance, fully crossed matched-scaffold interventions with a neutral placebo, and held-out-model-family prediction. The three profiles are descriptive; inferential analyses use the model sets specified for each analysis.

2 Related Work

Cognitive theory and models. The Cattell-Horn-Carroll taxonomy (McGrew 2009) and the unity-diversity model of executive functions (Miyake et al. 2000) motivate our working-memory, cognitive-control, and episodic-memory groupings. Theory of mind and metacognition draw on false-belief and metacognitive-monitoring traditions (Wellman, Cross, and Watson 2001; Lichtenstein and Fischhoff 1977; Persaud, McLeod, and Cowey 2007). WMF-AM (Hou et al. 2026) provides a depth-parameterized cumulative-state-tracking probe associated with downstream agent performance. Whether one axis suffices, or per-ability profiles add signal, is the separability question CogArena tests.

Cognitive evaluation of LLMs. CogBench (Coda-Forno et al. 2024) derives ten behavioral metrics from seven cognitive experiments, fits multilevel models, and studies prompt effects. It characterizes task-specific behavior; CogArena instead asks whether scores from multiple paradigms support a family-general grouping structure. Capacity-specific work studies executive function (de Langis et al. 2026), but not a multi-construct taxonomy. Reviewing 445 LLM benchmarks, Bean et al. (2025) find construct validity largely absent; Jung et al. (2026) likewise show that psychometric reliability need not imply ecological validity.

Intervention and scaffold validity. Theory-guided prompting can elicit latent task performance. NeuReasoner maps where a modular cognitive elicitation procedure helps on CogBench and conventional reasoning benchmarks (Javadov et al. 2026). Yet direct prompt gains confound useful semantic content with scaffold format and auxiliary context. He et al. (2026) address this using format-only, misleading, and wrong-fact controls for mathematical concept scaffolds. CogArena asks a different, discriminant question. Every targeted scaffold is crossed with every theory grouping and compared with a length-matched neutral placebo, so generic improvement cannot satisfy the diagonal estimand. This is an intervention on instructions and scores, not on an internal cognitive mechanism.

The structure of LLM abilities. A positive manifold and a dominant general factor are well documented across CHC-classified benchmark scores (Ilić and Gignac 2024), benchmark batteries (Kipnis et al. 2025), and a neuropsychological battery with one task per dimension (Haznitrama, Ardi, and Oh 2026). Burnell et al. (2023) recover correlated factors, but achievement factors need not validate theory-defined cognitive constructs. Ability-scale instruments (Zhou et al. 2026) typically assume their dimensions. In our comparison, CogArena alone combines within-paradigm construct checks, convergent and discriminant analysis, fully crossed scaffold-specificity tests, and held-out-model-family prediction for

the same taxonomy (Appendix Table S12).

3 The CogArena Benchmark

Figure 1 summarizes the 13 paradigms, their 5 theory-motivated groupings, and the validation workflow. Each paradigm is adapted from a validated human experiment with published reference data, although the original endpoint is often reaction time, span, or calibration rather than accuracy. Here a "paradigm" is a standardized experimental task with a fixed procedure and an expected behavioral effect, not a modeling approach. The groupings follow established cognitive-science taxonomies, but Section 5.2 tests rather than assumes that they form stable dimensions.

Accuracy is the common profile endpoint, while construct-relevant contrasts are evaluated separately in Section 5.1. Deterministic item- or episode-level scoring yields one model-level accuracy per paradigm, and the resulting 13 scores form the model-by-paradigm matrix analyzed in Section 5.2. Appendix Table S13 provides full paradigm definitions, source anchors, human sample sizes, adaptation ratings, and evaluation modes; Appendix Table S11 reports the intended signals and observed signature evidence.

3.1 Construction and Construct Checks

Procedural Generation. All task items are procedurally generated. Each generator randomizes surface content (names, objects, word lists) and sets each item's condition and difficulty by design. This mitigates data contamination from training corpora (probed directly in Section 5). Static benchmark items are known to inflate measured ability relative to freshly generated variants of the same problems (Mirzadeh et al. 2025). The main battery excludes canonical stimuli such as the Sally-Anne scenario; classic items are used only in the separate contamination probe. Appendix S1.1 gives one generated example per paradigm.

Adaptation Distance. We rate each paradigm's adaptation distance from the original human experiment (Low/ Medium/High). Language-mediated tasks (false belief, DRM, metacognition) preserve the core construct well (Low). Tasks depending on perceptual-motor processing (Stroop color naming, Go/No-Go inhibition) require more adaptation (Medium). Paradigms whose original form is fundamentally non-linguistic (High distance, e.g. mental rotation) cannot be faithfully text-adapted and are excluded. CogArena retains only Low- and Medium-distance paradigms. These ratings are author judgments, not a computed metric. The behavioral-signature checks below provide the empirical test of whether each adapted paradigm still reproduces the expected human directional effect.

Two-Level Validation Framework. We assess validity at the paradigm and profile levels. At the paradigm level, three complementary diagnostics audit the text adaptations where the design permits.

1. **Behavioral signatures.** Does the expected directional effect hold? (e.g., congruent $>$ incongruent for Stroop)
2. **Difficulty gradients.** Does accuracy decline as construct-relevant demand increases? (e.g., more sources $\rightarrow$ lower accuracy; clearest for source monitoring and n-back load)
3. **Cross-modal checks.** Do text and image versions produce different patterns? (for the three paradigms with a visual form)

The behavioral-signature analysis is the primary adaptation check, with per-paradigm outcomes reported in Section 5.1. At the profile level, we test whether the proposed groupings show convergent and discriminant separation, respond selectively to matched scaffolds, and improve prediction for held-out model families. Family-clustered intervals quantify uncertainty. Post-hoc robustness analyses include within-family centering and construct-native rescoring.

3.2 Evaluation Modes

Text evaluation constitutes the primary benchmark. VLM evaluation is a targeted cross-modal adaptation check on three paradigms, while agent evaluation is an exploratory pilot with four models. All three use a shared Gymnasium-style interface, a single reset/step contract whose observation space and episode length vary by mode. Full API and environment specifications appear in the appendix.

- **Text LLM.** All 13 paradigms use text prompts with paradigm-specific scoring. Ten use single-response evaluation; n-back, operation span, and CVLT use multi-turn episodes (Section 4).
- **VLM.** Image stimuli for Stroop, Flanker, and false belief replace text descriptions.
- **Agent pilot.** N-back and false belief from the battery, plus the Wisconsin Card Sorting Test, use multi-turn tool access for memory, calculation, and note-taking.

4 Experimental Setup

Model Pools. We evaluate 20 open-weight text LLMs from nine families, spanning 0.5B–47B parameters. This primary pool is used for per-paradigm accuracy, behavioral-signature, and cross-modal analyses. Dimensional-structure and scaling-robustness analyses additionally use an expanded pool of 55 models from over 20 families. Open checkpoints provide known parameter counts and family lineage while enabling reproducible local evaluation. We also evaluate six VLMs on three image-based paradigms and four text LLMs in the agent pilot. Full checkpoint lists are provided in Table S1.

Items and Administration. All models receive the same procedurally generated items. Most paradigms contain 50 items across three designed difficulty levels; Stroop and Flanker contain 66 items spanning congruent and incongruent conditions. Ten paradigms use single-turn evaluation, whereas n-back, operation span, and CVLT are administered as multi-turn episodes. Multi-turn prompts retain the initial instructions and a sliding window of the most recent 30 transcript lines. Exact manifests, seeds, and evaluation counts are reported in the appendix.

Scoring and Serving. Scoring is deterministic and rule-based, without an LLM judge. Single-answer items use normalized exact or regular-expression matching, multi-part items permit partial credit, and the primary cross-paradigm outcome is answer accuracy. Operation span and CVLT use recall-based scorers, with an alternative operation-span parser reported as a specification sensitivity. Final analyses use corrected scorers and regenerated affected items; Appendix S1.2 reports the correction scope and scoring sensitivities. Models are served locally through Ollama using default quantization and greedy decoding.

Intervention-Validity Study. After the observational study, we outcome-froze a fully crossed intervention protocol using 12 checkpoints from six families, all 13 paradigms, and 18 held-out items per paradigm. Seven conditions comprise baseline, a length-matched neutral placebo, and five answer-free scaffolds targeting the five proposed cognitive groupings; every scaffold is applied to every paradigm. For scaffold s, selectivity S_s is its placebo-adjusted gain on matched paradigms minus its gain on nonmatched paradigms, and Γ is the equal-weight mean across scaffolds. Label permutations test diagonal alignment, family-by-item resampling quantifies uncertainty, an exact sign-flip test assesses cross-family consistency, and leave-one-family-out prediction tests transport. Confirmation requires all nine frozen gates to pass. The protocol was frozen before formal outcome inspection but was not preregistered; full prompts and decision rules appear in Appendix S1.11.

Intervention Panel and Resampling. The crossed study contains 19,656 model-item-condition records. Its panel includes two checkpoints from each of Qwen2.5, Gemma2, Llama2, Gemma3, Falcon3, and OLMo2. The exact mapping test enumerates all 120 scaffold-to-group assignments. The crossed interval resamples the six families and the 18 items per paradigm while preserving condition pairing within each item.

Scaffold Contents. The five answer-free scaffolds provide a working-memory ledger, rule rehearsal, source binding, belief-state ledger, or metacognitive forecast. They specify how to organize a response without supplying item answers. Applying every scaffold to all 13 paradigms separates matched-grouping selectivity from generic prompting benefits in the off-target cells.

5 Results

We first assess whether individual paradigm adaptations preserve their expected behavioral signatures. We then test whether the proposed groupings separate in model scores, improve prediction for held-out families, and respond selectively to matched scaffolds. Scaling and auxiliary checks provide secondary evidence.

5.1 Paradigm-Level Construct Validity

Aggregate directional effects hold for most paradigms, but checkpoint-level replication is mixed. Under one-sided checkpoint binomial tests with BH correction, DRM false memory (18/20 models), Flanker (18/20), and n-back load (15/20) replicate; false belief is directionally consistent but nonsignificant (12/20), and text Stroop does not replicate (7/20). Treating merged model families as the sampling units retains directional evidence for Flanker (10/10 families, $p_{\mathrm{BH}} = .0049$) and DRM (9/10, $p_{\mathrm{BH}} = .027$), but not n-back (7/10, $p_{\mathrm{BH}} = .215$). EPITOME's forced-choice rerun reproduces the expected desire-over-belief ordering in 25/35 expansion models ($p = .008$) and 19/21 merged families ($p = .0001$). Thus construct labels are credible for some paradigms but not licensed uniformly by provenance alone.

Figure 2 makes the mixed adaptation evidence explicit. Strong Flanker and DRM contrasts coexist with weaker Stroop and false-belief signatures. This heterogeneity motivates the profile-level tests below.

Two paradigm-specific constraints require particular caution. Go/No-Go contains 42 GO trials among 50, so an all-GO responder scores 84% without following the rule. Recall-scored CVLT retains the studied list in the running transcript, making textual availability part of the construct.

5.2 Dimensional Structure of Model Performance

Across 55 models, 77 of 78 paradigm correlations are positive. The first principal component explains 49.8% of paradigm-score variance and correlates at $r = .99$ with mean accuracy, indicating a broad performance axis. Within-grouping correlations average .496 and cross-grouping correlations .415, a modest difference under the primary scorer ($\delta = .081$, exact two-sided $p = .057$). The canonical sensitivity is similar (Table 1).

Family-aware analyses weaken the distinction. The merged-family interval includes zero (95% CI $[-.012, .145]$); within-family centering gives $\delta = .011$ ($p = .798$), and 24 family centroids give $\delta = .079$ ($p = .184$). These estimates show a grouping advantage, but not stable family-general dimensions. Joint family-item analyses appear in Appendices S1.5–S1.7.

Construct-native scoring reverses the raw contrast ($\delta = -.02$, $p = .76$) and reduces the first-component share to about 40%. Row-mean residualization also remains null ($\delta = .03$, $p = .68$). The seven alternative endpoints retain split-half reliabilities of .65–.99. Residualization, difficulty, and range checks preserve some positive estimates but do not resolve their family and scoring dependence (Appendices S1.5–S1.7).

Simulation Calibration. Calibrated simulations characterize the structure test's operating properties. At a group-factor arm with a .15 within-grouping correlation increment, the raw test detects structure in 92% of repetitions, with realized δ averaging .11. Across 1,000 general-factor-only matrices, row-mean residualization has type-I rates of .026–.031 and PC1 removal gives .054–.061. Horn parallel analysis retains one component for accuracy scores. It retains two for construct-native scores, but the second separates difference and signal-detection endpoints from recall and accuracy endpoints across grouping boundaries. The extra component therefore resembles a scoring-method factor rather than the proposed taxonomy.

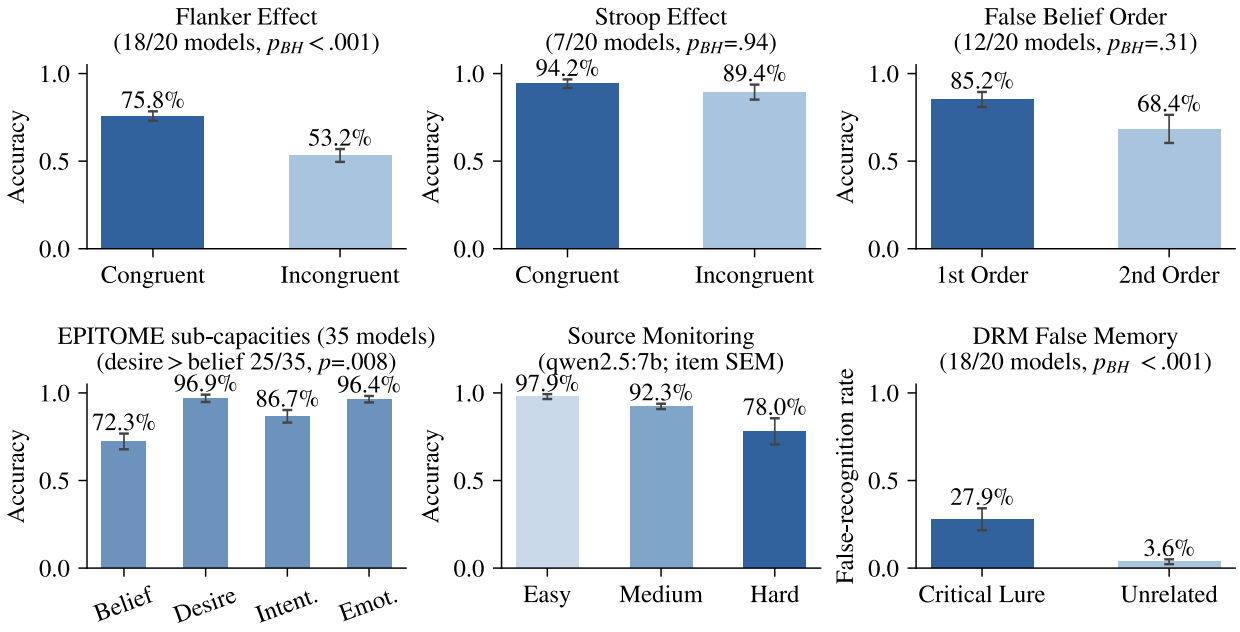

Figure 2: Paradigm-level construct diagnostics. Bars show corrected mean accuracy, except that DRM shows false-recognition rate. Error bars are standard errors across models, except for the source-monitoring item sweep for Qwen2.5-7B. Titles report checkpoint-level directional counts and BH-adjusted tests where applicable. Strong Flanker and DRM contrasts coexist with weaker Stroop and false-belief signatures, motivating profile-level validation rather than assuming validity from paradigm labels.

Analysis view	δ	p_2	Family 95% CI
Strict accuracy	.081	.057	$[-.012, .145]$
Canonical accuracy	.087	.042	$[-.005, .156]$
Within-family centered	.011	.798	$[-.087, .085]$
Family centroids	.079	.184	not est.
Construct-native	$-.020$	.760	$[-.110, .050]$

Table 1: Dimensional-separation estimates across scoring and family views. The primary strict estimate is small, and its inferential status changes across defensible views.

Where Grouping Structure Strengthens. Two post-hoc views yield larger positive estimates. Across 11 paradigms with designed difficulty tiers, δ rises from .117 on easy items to .140 on medium and .169 on hard items. Merged-family intervals exclude zero at every tier but include .15. Jointly excluding text Stroop, Go/No-Go, and CVLT increases accuracy separation to $\delta = .147$ ($p_2 = .021$), whereas construct-native separation remains $\delta = .095$ ($p = .441$). No family-clustered interval was computed for the joint deletion. These analyses recover grouping structure in restricted views, but do not establish scoring- and family-invariant dimensions.

5.3 Cross-Family Transport and Intervention Selectivity

Across 24 held-out model families, grouping labels do not improve target-paradigm RMSE beyond a general-component predictor (relative gain -1.8%, family-bootstrap CI $[-6.3\%, 2.0\%]$), and only 3 of 13 target paradigms improve. Construct-native scores likewise fail to transport (relative gain -4.73%, 95% CI $[-6.13\%, -2.87\%]$; 0/13 improve). Adjacent-administration model-centered profile cells are nevertheless stable across eight eligible paradigms (ICC=.979), making random replay variation an unlikely explanation for the transport null.

Relative to the neutral placebo, the five targeted scaffolds produce a small aggregate diagonal advantage ($\Gamma = .0199$, crossed family-by-item 95% CI $[.0041, .0360]$). The exact two-sided family sign-flip test gives $p = .063$, and no scaffold-specific mapping contrast survives BH correction. These results indicate a weak battery-level alignment tendency rather than robust scaffold-specific effects.

The frozen all-nine rule fails (Figure 4). The gates jointly require a positive crossed interval, consistent family direction, correct scaffold-grouping alignment, low protocol-invalid rates, robustness to invalid, empty, and unparseable responses, stability after response-length adjustment, and improved held-out-family prediction. Six pass. Predictive transport fails, and the empty-response and operation-span parse exclusions leave some cells below the frozen minimum.

Consistent with the observational transport result above, selective intervention-by-group terms do not improve leave-one-family-out prediction ($\Delta LL = -.904$; 2/6 families improve). The study therefore shows weak in-battery alignment without held-out-family confirmation.

The three intervention tests separate assignment, family consistency, and transport. The intended scaffold-grouping mapping outperforms alternative assignments, consistency

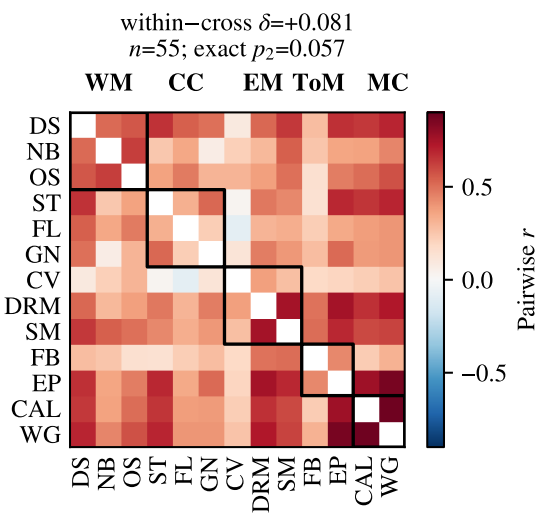

Figure 3: Pearson correlations among corrected paradigm accuracies across 55 models, ordered by the five proposed groupings. Black outlines mark within-grouping blocks, and white diagonal cells omit self-correlations. Abbreviations follow the paradigm inventory in Appendix Table S13, with NB for n-back, OS for operation span, CV for CVLT, and CAL for confidence calibration. A separable taxonomy would produce consistently higher correlations inside the outlined blocks. Instead, correlations are predominantly positive across the matrix, and several of the strongest cross grouping boundaries.

Wording	Γ	95% CI	Families +	LOFO ΔLL
Frozen	.0199	[.0041,.0360]	5/6	−.904
Alternate	.0134	[−.0030,.0298]	4/6	+.771

Table 2: Scaffold-wording comparison under the same design. Exact mapping $p_2 = .0167$ and .0333 for the frozen and alternate wordings; 2/6 and 3/6 held-out families improve. Both fail the complete rule.

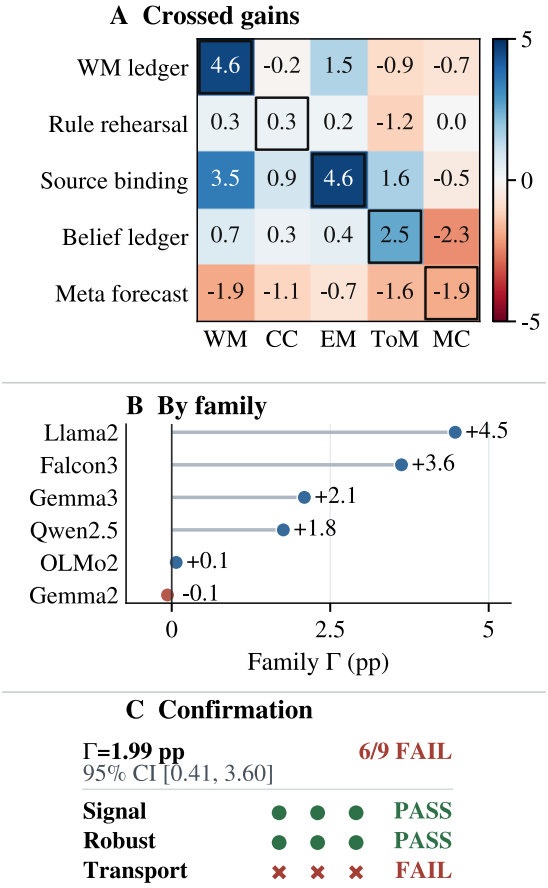

Figure 4: Intervention-validity evidence. (A) Target-minus-placebo gains in percentage points. Boxes mark matched scaffold-group pairs, with grouping abbreviations from Table 1. (B) Descriptive family-level Γ estimates. (C) Nine frozen gates grouped by signal, robustness, and transport. Circles pass and crosses fail. The positive aggregate tendency does not satisfy transport, so the all-gates decision is FAIL. Full criteria appear in Appendix S1.11.

across six families remains borderline, and prediction to an unseen family fails.

A post-hoc alternate-wording replication retains a smaller positive diagonal estimate, but its interval includes zero and the all-nine rule again fails (Table 2). Because it reuses the same models and held-out items, this comparison isolates wording sensitivity rather than providing an independent replication.

Replacing the placebo with the no-scaffold baseline preserves the diagonal tendency (Γ = .0207, 95% CI [.0034,.0382], exact mapping $p = .0167$). The group-differential placebo contribution is near zero, so selective placebo harm does not explain the alignment. Targeted arms nevertheless average 0.81 percentage points below baseline, separating selective alignment from general improvement. Full audits appear in Appendix S1.11.

Across 13 post-hoc leave-one-paradigm-out analyses, Γ remains positive at .0164–.0237. A three-level bootstrap over families, paradigms, and items gives a 95% CI of [.0007,.0420]. Because each grouping contains only two or three paradigms, this supports alignment within the finite battery rather than a population claim over possible paradigms.

Together, the covariance, transport, and intervention results support the groupings as an organizing taxonomy, but not as stable family-general dimensions.

5.4 Scaling and Auxiliary Validity Checks

Scaling is paradigm-dependent, with correlations with log parameter count ranging from .12 to .74; the heterogeneous ordering persists in the expanded and family-aware analyses (Appendix S1.4). Representative single-response accuracies and complete 20-model multi-turn accuracies are reported in Tables S14 and S2.

Validation level	Main result
Paradigm checks	Flanker and DRM replicate in 18/20 checkpoints; EPITOME in 19/21 families. Stroop and false belief are weaker. Evidence is paradigm-specific.
Covariance	PC1 explains 49.8%; $\delta = .081$; the family interval includes zero. Broad competence dominates a modest grouping advantage.
Scaffold selectivity	$\Gamma = .0199$ and family $p = .063$; alternate wording is smaller. Alignment is weak rather than scaffold-specific.
Family transport	RMSE changes by -1.8% and $\Delta LL = -.904$. Grouping information does not help on unseen families.

Table 3: Evidence across four cumulative validation levels. Later failures limit the stronger grouping claim without erasing paradigm-level evidence.

Cross-modal evaluation shows that text adaptation can alter a construct. Five VLMs with consistently parseable Stroop labels recover the human-direction congruency contrast absent in text, while image false-belief accuracy ranges from 0% to 66%. These unpaired descriptive checks motivate adaptation audits rather than estimate a modality effect. Matched human accuracies exist only for false belief and EPITOME (Strachan et al. 2024; Jones, Trott, and Bergen 2024). Grouping scores correlate with three external benchmarks in 10 of 15 BH-corrected pairs, but the small samples make these exploratory. A contamination probe finds no correction-surviving classic-item advantage and cannot exclude small effects. Full results appear in Appendices S1.3 and S1.8–S1.12.

The four levels answer progressively stronger questions. A behavioral signature supports interpretation of one paradigm. Covariance asks whether proposed groupings cohere beyond broad performance. Scaffold specificity asks whether a matched manipulation shifts them selectively. Transport asks whether either observational grouping scores or intervention selectivity improves prediction for an unseen family. Failure at a later level limits the dimensional claim without erasing earlier paradigm-level evidence.

6 Discussion and Limitations

Text Adaptation Boundaries. Propositional paradigms such as DRM, wagering, and calibration are comparatively well preserved, and Flanker interference replicates in text. Automatic color-word conflict does not survive text Stroop, while text Go/No-Go reduces to explicit rule following with an exploitable base rate. Human sources therefore provide directional anchors rather than a common human-LLM scale.

Interpreting the Boundary Result. The covariance, intervention, and transport analyses distinguish a useful taxonomy from validated cognitive dimensions. Positive raw contrasts and matched-scaffold gains argue against claiming that grouping structure is absent. Yet the broad common axis, family-aware uncertainty, scoring sensitivity, and failed transport prevent treating grouping means as stable traits. The groupings remain useful for sampling and organization, but paradigms with replicated signatures are the best-supported reporting units. With one text modality and only two or three paradigms per grouping, the boundary result applies to this battery and model pool rather than LLM cognitive architecture in general.

Implications for Cognitive Benchmarking. CogBench and related batteries show that LLMs can reproduce informative task-level behavioral patterns (Coda-Forno et al. 2024). CogArena addresses the next measurement question, namely when scores from several paradigms warrant a shared cognitive label. That claim requires more than task coverage or correlated accuracy. The proposed grouping should survive construct checks, separate from other groupings under family-aware inference, respond selectively to a matched manipulation, and improve prediction for an unseen model family. Applying all four requirements to one taxonomy is the main methodological contribution. The result is useful even when confirmation fails because it distinguishes a descriptive benchmark organization from a validated profile of transportable dimensions.

Scope of the Intervention Evidence. The frozen study measures prompt-contingent score alignment using one wording per target. A post-hoc alternative preserves the direction but reuses the same models and items. The neutral placebo controls prompt presence and approximate length; misleading and wrong-content controls remain future work (He et al. 2026). The panel has six families and only 2–3 paradigms per grouping.

Limitations. (1) only open-weight models (to 72B dense, one 141B-total MoE), no closed/frontier; (2) agent evaluation is pilot-scale ($n = 4$); (3) contamination is tested on single-turn paradigms only; (4) VLM covers 3 paradigms; (5) matched human accuracies exist for only 2/13 paradigms (Strachan et al. 2024; Jones, Trott, and Bergen 2024); (6) memory scaling is scorer-dependent and CVLT measures availability as much as retention; (7) external benchmark correlations pair default-quantized CogArena scores with published full-precision scores; (8) shared metacognition items, Go/No-Go's base rate, and 2–3 paradigms per grouping limit structural inference; and (9) the wording replication reuses the same six families and held-out items. All conclusions are properties of this text battery, not claims about LLM cognitive architecture in general.

7 Conclusion

CogArena provides a reusable framework for deciding when adapted cognitive-benchmark scores warrant dimensional labels. Across 13 paradigms and 55 models, broad competence dominates, while grouping structure is modest and family- and scoring-dependent. Matched scaffolds show a small tendency, but confirmation and held-out-family prediction fail. The five groupings therefore remain an organizing taxonomy, not established latent abilities; together, signatures, covariance, interventions, and transport provide a stricter basis for cognitive labels.

References

Baron-Cohen, S.; Leslie, A. M.; and Frith, U. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1): 37–46.

Bean, A. M.; Kearns, R. O.; Romanou, A.; et al. 2025. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. In *Advances in Neural Information Processing Systems 38 (NeurIPS), Datasets and Benchmarks Track*.

Binz, M.; Akata, E.; Bethge, M.; Brändle, F.; Callaway, F.; Coda-Forno, J.; et al. 2025. A foundation model to predict and capture human cognition. *Nature*, 644(8078): 1002–1009.

Binz, M.; and Schulz, E. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6): e2218523120.

Bugaud, Z. 2026. A Cognitive Battery for Foundation Models: Theory-Grounded Benchmarks for Attention, Learning, Metacognition, Executive Function, and Social Cognition. In *ICML 2026 Workshop on Combining Theory and Benchmarks*.

Burnell, R.; Hao, H.; Conway, A. R. A.; and Hernández-Orallo, J. 2023. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062*.

Coda-Forno, J.; Binz, M.; Wang, J. X.; and Schulz, E. 2024. CogBench: A large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Contreras, J. M. 2026. An LLM-Native Psychometric Instrument Does Not Predict LLM Behavior: Evidence Across 25 Models. *arXiv preprint arXiv:2606.09843*.

de Langis, K.; Park, J. I.; Hu, B.; Le, K. C.; Schramm, A.; Mensink, M. C.; Elfenbein, A.; and Kang, D. 2026. Strong Memory, Weak Control: An Empirical Study of Executive Functioning in LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5971–5986.

Delis, D. C.; Kramer, J. H.; Kaplan, E.; and Ober, B. A. 2000. *California Verbal Learning Test–Second Edition (CVLT-II): Adult Version Manual*.

Eriksen, B. A.; and Eriksen, C. W. 1974. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16(1): 143–149.

Fischhoff, B.; Slovic, P.; and Lichtenstein, S. 1977. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 3(4): 552–564.

Haznitrama, F. G.; Ardi, F. R.; and Oh, A. 2026. A Neuropsychologically Grounded Evaluation of LLM Cognitive Abilities. *arXiv preprint arXiv:2603.02540*.

He, J.; Dai, S.; Qiao, X.; Li, J.; Yan, Y.; and Hu, X. 2026. Beyond Direct Gains: Matched Controls for Evaluating Concept Scaffolds. In *ICML 2026 AI4Math Workshop*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Hou, D.; Jiang, L.; Li, D.; Li, Z.; Lin, F.; and Yamada, K. D. 2026. WMF-AM: Probing LLM Working Memory via Depth-Parameterized Cumulative State Tracking. *arXiv preprint arXiv:2603.27343*.

Ilić, D.; and Gignac, G. E. 2024. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106: 101858.

Javadov, A.; Aitkazinov, S.; Hoesli, T.; von Wangenheim, F.; Schuller, B.; and Ollier, J. 2026. NeuReasoner: Theory-Grounded Mapping of Reasoning Elicitation Boundaries. *arXiv preprint arXiv:2606.29971*.

Johnson, M. K.; Hashtroudi, S.; and Lindsay, D. S. 1993. Source monitoring. *Psychological Bulletin*, 114(1): 3–28.

Jones, C. R.; Trott, S.; and Bergen, B. 2024. Comparing Humans and Large Language Models on an Experimental Protocol Inventory for Theory of Mind Evaluation (EPITOME). *Transactions of the Association for Computational Linguistics*, 12: 803–819.

Jung, J.; Lutz, M.; Sen, I.; and Strohmaier, M. 2026. Do Psychometric Tests Work for Large Language Models? Evaluation of Tests on Sexism, Racism, and Morality. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8143–8173. Association for Computational Linguistics.

Kipnis, A.; Voudouris, K.; Schulze Buschoff, L. M.; and Schulz, E. 2025. metabench: A Sparse Benchmark of Reasoning and Knowledge in Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Lichtenstein, S.; and Fischhoff, B. 1977. Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20(2): 159–183.

MacLeod, C. M. 1991. Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109(2): 163–203.

McGrew, K. S. 2009. CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37(1): 1–10.

Mirzadeh, I.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; and Farajtabar, M. 2025. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*.

Miyake, A.; Friedman, N. P.; Emerson, M. J.; Witzki, A. H.; Howerter, A.; and Wager, T. D. 2000. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, 41(1): 49–100.

Momentè, F.; Suglia, A.; Giulianelli, M.; Ferrari, A.; Koller, A.; Lemon, O.; Schlangen, D.; Fernández, R.; and Bernardi, R. 2025. Triangulating LLM Progress through Benchmarks, Games, and Cognitive Tests. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 20051–20072. Association for Computational Linguistics.

Pelegrina, S.; Lechuga, M. T.; García-Madruga, J. A.; Elosúa, M. R.; Macizo, P.; Carreiras, M.; Fuentes, L. J.; and Bajo, M. T. 2015. Normative Data on the N-Back Task for Children and Young Adolescents. *Frontiers in Psychology*, 6: 1544.

Persaud, N.; McLeod, P.; and Cowey, A. 2007. Post-decision wagering objectively measures awareness. *Nature Neuroscience*, 10(2): 257–261.

Redick, T. S.; Broadway, J. M.; Meier, M. E.; Kuriakose, P. S.; Unsworth, N.; Kane, M. J.; and Engle, R. W. 2012. Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, 28(3): 164–171.

Roediger, H. L.; and McDermott, K. B. 1995. Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4): 803–814.

Serapio-García, G.; Safdari, M.; Crepy, C.; Sun, L.; Fitz, S.; Romero, P.; Abdulhai, M.; Faust, A.; and Matarić, M. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12): 1954–1968.

Strachan, J. W. A.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; Graziano, M. S. A.; and Becchio, C. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8: 1285–1295.

Stroop, J. R. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6): 643–662.

Trott, S.; Rivière, P. D.; and Jones, C. R. 2026. Do Different Theory of Mind Tasks for LLMs Measure the Same Thing? In *ACL 2026 Workshop on Evaluating Evaluations (EvalEval)*.

Van der Elst, W.; Van Boxtel, M. P. J.; Van Breukelen, G. J. P.; and Jolles, J. 2006. The Stroop Color-Word Test: Influence of age, sex, and education; and normative data for a large sample across the adult age range. *Assessment*, 13(1): 62–79.

Votruba, K. L.; and Langenecker, S. A. 2013. Factor structure, construct validity, and age- and education-based normative data for the Parametric Go/No-Go Test. *Journal of Clinical and Experimental Neuropsychology*, 35(2): 132–146.

Wechsler, D. 2008. *WAIS-IV Administration and Scoring Manual*.

Wellman, H. M.; Cross, D.; and Watson, J. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3): 655–684.

Yang, Y.; Miao, C.; Li, W.; and Wu, Y. 2026. ActTrait-Bench: Quantifying the Knowledge–Decision Gap in Large Language Models via Human-Grounded Behavioral Validation. *arXiv preprint arXiv:2605.29791*.

Zhou, L.; Pacchiardi, L.; Martínez-Plumed, F.; Collins, K. M.; et al. 2026. General scales unlock AI evaluation with explanatory and predictive power. *Nature*, 652: 58–67.

S1 Additional Results and Details

S1.1 Models Evaluated

Table S1 lists all 55 open-weight text LLMs by exact Ollama registry tag, family, and parameter count, marking the 20 that constitute the full 13-paradigm battery; all 55 enter the scaling and convergent and discriminant validity analyses. The six-VLM cross-modal subset and four agent configurations are listed with their respective results in Sections S1.8 and S3.

S1.2 Example Items

One procedurally generated item per paradigm (seed = 42; novel stimuli, no contamination probes; long study lists abridged as [. . .]). "Expected" is the scorer's gold target; three entries also show the actual qwen2.5:7B answer.

Digit Span. "Repeat the digit sequence in the SAME order. Digits: 3 6 2." **Expected answer.** 3 6 2.

N-Back (n=2). "For each of 24 tokens, respond MATCH if it equals the token 2 positions earlier, else NO MATCH. First token: KW [. . .]" **Expected answer.** The per-token MATCH/NO-MATCH sequence (8 matches among 24).

Operation Span (set size 3). "For each item, verify an equation (YES/NO) then remember a letter; after all items recall the letters in order. Item 1: Is $(2 \times 9) - 6 = 12$? Remember: C [. . .]" **Expected answer.** C Z N.

Stroop. "The word "ONE" appears 7 times: ONE ONE ONE ONE ONE ONE ONE. How many times does the word appear?" **Expected answer.** 7. **qwen2.5:7B answer.** 7 (correct; counts despite the conflicting word meaning).

Flanker. "Stimulus: K K K S K K K. What is the CENTER letter?" **Expected answer.** S.

Go/No-Go. "Respond GO if the word is clothing, NO-GO if furniture. Trial 1: shorts." **Expected answer.** GO.

CVLT Word List. "Study a 14-word list over 5 trials, recalling after each: pilot, janitor, lawyer, [. . .], welder; then an interference list, then recall the original." **Expected answer.** The 14 studied words.

DRM False Memory. "Study themed lists (e.g. coat, arctic, polar, blizzard, shiver, winter, snow, freeze, chilly, ice [. . .]); then mark each test word OLD or NEW." The semantically central lure "cold" is never presented. **Expected answer.** The word "cold" should be marked NEW (models frequently false-alarm OLD).

Source Monitoring. "20 statements, each attributed to one of four similar speakers (Dr. Muller, Dr. Tanaka, Dr. Sullivan, Professor Sullivan); then identify who said each." **Expected answer.** For example, "goulash requires saffron" $\rightarrow$ Dr. Muller.

False Belief. "Astrid places a gold coin in the tote bag and leaves; Rafael then moves it to another container; where will Astrid look for it first?" **Expected answer.** the tote bag. **qwen2.5:7B answer.** the tote bag (correct).

EPITOME (ToM). "Nadia heard from Jia that the store is closed; actually it is open and Jia was mistaken. Does Nadia believe the store is open or closed? (A) open (B) closed." **Expected answer.** B.

Confidence Calibration. "How many flats are in the key of B-flat major? Give your answer and your confidence (0–

Model (Ollama tag)	Family	B	Bat.	Model (Ollama tag)	Family	B	Bat.
aya:8b	aya	8		olmo2:13b	olmo2	13	
command-r:35b	command-r	35	•	openchat:7b	openchat	7	
deepseek-llm:7b	deepseek	7		phi3:3.8b	phi	3.8	
deepseek-r1:7b	deepseek-r1	7	•	phi3:14b	phi	14	•
deepseek-r1:14b	deepseek-r1	14	•	phi4:14b	phi	14	
exaone3.5:7.8b	exaone	7.8		qwen2.5:0.5b	qwen2.5	0.5	•
falcon3:7b	falcon3	7		qwen2.5:1.5b	qwen2.5	1.5	•
falcon3:10b	falcon3	10		qwen2.5:3b	qwen2.5	3	•
gemma2:2b	gemma2	2	•	qwen2.5:7b	qwen2.5	7	•
gemma2:9b	gemma2	9	•	qwen2.5:14b	qwen2.5	14	•
gemma2:27b	gemma2	27	•	qwen2.5:32b	qwen2.5	32	•
gemma3:1b	gemma3	1		qwen2.5:72b	qwen2.5	72	
gemma3:12b	gemma3	12		qwen3:0.6b	qwen3	0.6	
gemma3:27b	gemma3	27		qwen3:1.7b	qwen3	1.7	
glm4:9b	glm4	9		qwen3:4b	qwen3	4	
internlm2:7b	internlm2	7		qwen3:8b	qwen3	8	
llama2:7b	llama2	7		qwen3:14b	qwen3	14	
llama2:13b	llama2	13		smollm2:360m	smollm2	0.36	
llama3.1:8b	llama3.1	8	•	smollm2:1.7b	smollm2	1.7	
llama3.1:70b	llama3.1	70		solar:10.7b	solar	10.7	
llama3.2:1b	llama3.2	1	•	stablelm2:1.6b	stablelm2	1.6	
llama3.2:3b	llama3.2	3	•	starling-lm:7b	starling	7	
mistral:7b	mistral	7	•	tinyllama:1.1b	tinyllama	1.1	•
mistral-nemo:12b	mistral	12		yi:6b	yi	6	
mixtral:8x7b	mixtral	47	•	yi:9b	yi	9	
mixtral:8x22b	mixtral	141		yi:34b	yi	34	•
nemotron-mini:4b	nemotron	4		zephyr:7b	zephyr	7	
olmo2:7b	olmo2	7					

Table S1: Complete list of the 55 open-weight text LLMs evaluated (exact Ollama registry tags). • marks the 20 models in the full 13-paradigm battery; all 55 are used in the scaling and convergent and discriminant validity (separability) analyses. B = parameters in billions (Mixtral entries are total parameters). All models are served through Ollama at each tag's default quantization (4-bit `Q4_K_M` for the 7B-class checkpoints); see the serving-configuration note.

100%)." **Expected answer.** 2. **qwen2.5:7B answer.** "Answer: 1, Confidence: 100%" (confidently wrong, a calibration failure).

Post-Decision Wagering. "What enzyme breaks down starch in saliva? Give your answer and whether you BET 10 points it is correct (YES: $\pm$10; NO: $+$2)." **Expected answer.** amylase.

Behavioral signatures.

- **Stroop.** Aggregate congruent performance (94.2%) exceeds incongruent performance (89.4%), and the ordering holds for 7/20 models. The small gap (4.8%) reflects weak text-based conflict; in humans, interference is primarily in RT (MacLeod 1991).
- **Flanker.** Aggregate congruent performance (75.8%) exceeds incongruent performance (53.2%), and the ordering holds for 18/20 models. The AI effect (+22.6%) is larger than the human effect ($\sim$4–5%), as text symbol parsing is harder than visual arrow identification.
- **False Belief.** Aggregate first-order performance (85.2%) exceeds second-order performance (68.4%), and the ordering holds for 12/20 models. This matches the human pattern in which second-order reasoning is consistently harder (Wellman, Cross, and Watson 2001).
- **EPITOME.** The 35-model expansion pool, whose per-item records support the sub-capacity split, follows the ordering desire (96.9%) $>$ emotion (96.4%) $>$ intention (86.7%) $>$ belief (72.3%). Belief tracking is the hardest sub-capacity, and the desire $>$ belief ordering replicates in 25/35 models.
- **Source Monitoring.** Accuracy degrades with difficulty, with easy (98%) $>$ medium (92%) $>$ hard (78%) for the representative qwen2.5:7b difficulty series. Difficulty levels correspond to increasing numbers of sources, matching the direction of the human pattern (Johnson, Hashtroudi, and Lindsay 1993).
- **DRM.** Models show the human false-memory effect, falsely recognizing the non-presented critical lure (27.9%) far more than unrelated words (3.6%); the effect replicates in 18/20 models, consistent with spreading-activation accounts (Roediger and McDermott 1995). Non-replicating small models discriminate at chance ($d' \approx 0$), so their absence of false memory reflects failure to encode the list rather than resistance to the illusion.
- **N-Back.** Accuracy decreases from 1-back (63.3%) to 2-back (54.4%), matching the expected load effect, but does not decrease further at 3-back (55.4%). The 2-back to 3-back plateau may reflect a floor effect or a qualitative shift in strategy at higher loads. Under the strict scorer, the mean is 56.8% with realistic variance (0–81%), revealing

Model	Size	NB	OS	CV
TinyLlama	1.1B	0	2	50
Qwen2.5	0.5B	73	19	78
Llama3.2	1B	48	47	93
Qwen2.5	1.5B	50	85	88
Gemma2	2B	58	89	99
Qwen2.5	3B	69	92	92
Llama3.2	3B	69	99	84
Qwen2.5	7B	74	67	**100**
Mistral	7B	67	45	**100**
DeepSeek-R1	7B	36	70	55
Llama3.1	8B	71	94	99
Gemma2	9B	78	**100**	87
Qwen2.5	14B	**81**	**100**	**100**
Phi3	14B	9	35	**100**
DeepSeek-R1	14B	40	93	86
Gemma2	27B	78	**100**	**100**
Qwen2.5	32B	80	99	**100**
Mixtral	47B	9	52	97
Yi	34B	72	53	**100**
Command-R	35B	76	41	**100**

The means are NB=56.8%, OS=69.2%, and CV=90.4%. The corrected scorers use exact match for n-back, serial-position recall credit for OS, and recall-based list scoring for CV.

Table S2: Multi-turn paradigm accuracy (%) for all 20 models. NB = N-Back, OS = Operation Span, CV = CVLT Word List.

that n-back is a capacity-limited task where even large models do not reach ceiling. Phi3-14B (9%) and Mixtral-47B (9%) show near-floor performance despite markedly higher operation span, suggesting a dissociation within working memory.

- **Operation Span.** Mean 69.2% under serial-position recall credit shows working memory under dual-task demand is not at ceiling for most models, with a 2–100% range providing strong discriminative power. TinyLlama sits at the floor (2%), three models reach 100%, and DeepSeek-R1-14B (93%) clearly exceeds DeepSeek-R1-7B (70%).

The six behavioral-signature and difficulty diagnostics are visualized in the main text. The continuous effects and family-level sensitivity reported here provide the supporting detail.

Family-level signature sensitivity. The checkpoint binomial tests above can overstate precision when several checkpoints share a model lineage. We therefore averaged each directional contrast within family and repeated the one-sided exact direction test with family means as the sampling units. Under the merged family labels used by the main family-aware analysis, Flanker remains positive in 10/10 families ($p_{\mathrm{BH}} = .0049$) and DRM in 9/10 ($p_{\mathrm{BH}} = .027$). N-back is positive in 7/10 families but does not survive this sensitivity ($p_{\mathrm{BH}} = .215$); false belief is also 7/10 ($p_{\mathrm{BH}} = .215$), and Stroop is 5/10 ($p_{\mathrm{BH}} = .623$). The separately evaluated EPITOME expansion is positive in 19/21 merged families ($p = .0001$). Continuous family-mean sign-flip tests and results under raw lineage labels are provided in the accompanying machine-readable artifact. We therefore use the checkpoint counts descriptively and treat the Flanker, DRM, and EPITOME patterns as the family-replicated signatures.

Multi-turn scoring contract. Operation-span recall uses serial-position credit against the target sequence. Final analyses use two deterministic parsers. The primary parser, frozen before final recomputation after reviewing production response formats, extracts the final explicit recall enumeration, accepts comma-, space-, and line-separated formats, and scores refusals, hedges, and non-enumerations as incorrect. A canonical whitespace-splitting parser is carried through every analysis as a scoring-specification sensitivity (headline δ=0.087, exact two-sided p=.042, versus δ=0.081, p=.057 under the primary parser). Reported numbers use no human adjudication. CVLT uses unique-hit capped recall against the studied list on production-designated recall turns. Duplicates count once, recall cannot exceed one, turns receive credit at recall $\geq .5$, and episode accuracy is their mean. Because the studied list remains visible in context, the resulting score measures availability as much as retention. N-back turns use the same strict parsing rules; unparseable turns count as errors for accuracy and are dropped from the construct-native d' recoding.

Stroop Across Text and Images. Text Stroop (92%) does not engage automatic color-word processing. In the image version, the five VLMs that consistently return parseable labels all show a human-direction accuracy congruency effect (MacLeod 1991). Qwen2.5-VL scores 100%/84% on congruent/incongruent trials (92% overall), MiniCPM-V 100%/96% (98%), Llama3.2-Vision 100%/82% (91%), Gemma3 100%/74% (87%), and LLaVA-7B 100%/0% (50%), a pattern consistent with reading the printed word rather than reporting its ink color. Moondream returns an empty completion on 85 of 100 trials; blanks are preserved and scored as incorrect (9% overall), so its score mainly reflects format failure.

False Belief Across Text and Images. Text false belief averages 77%. Image performance remains heterogeneous across the six VLMs. Qwen2.5-VL reaches 66%, MiniCPM-V 54%, Llama3.2-Vision 38%, Gemma3 10%, and LLaVA-7B and Moondream 0%. Each story is shown as a single four-panel montage. Final image scoring requires either the exact location label or a unique answer anchored to where the queried character will first look; free-form scene descriptions that merely mention the believed location do not score. Under this response contract, the result is a cross-modal adaptation check of visual belief attribution rather than pure theory-of-mind measurement.

S1.3 Performance Profiles

Figure S1 shows descriptive grouping-score summaries for the Qwen2.5 family. The 0.5B model is uniformly low, while larger checkpoints improve by different amounts across groupings. Cross-family comparisons reveal that Mistral-7B trails Qwen2.5-7B on false belief (68% vs. 100%) while nearly matching it on digit span (86% vs. 98%),

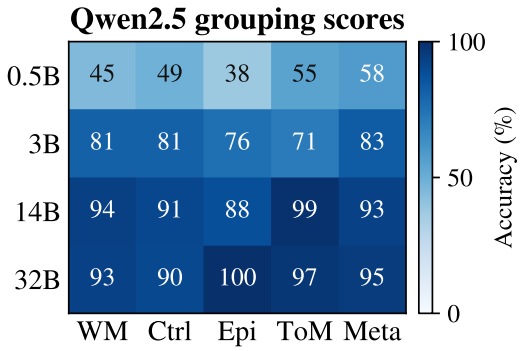

Figure S1: Descriptive Qwen2.5 grouping scores. Rows are checkpoints, columns are the five proposed groupings, and cells show mean accuracy as a percentage. The heatmap visualizes score heterogeneity without treating polygon area as meaningful.

and DeepSeek-R1-7B shows a pronounced descriptive per-paradigm dissociation.

Serving Configuration and Quantization. The observational and cross-modal batteries are served locally through Ollama's OpenAI-compatible endpoint (`/v1/chat/completions`) using the exact registry tags in Table S1. Requests use greedy decoding (`temperature`=0) and `max_tokens`=1024 (256 for VLM calls); seed 42 controls item generation and is not passed as a decoding seed. Bare tags use the registry default quantization at run time (typically `Q4_K_M` for 7B-class checkpoints). The study is tag-pinned because no immutable registry snapshot was captured. Fifty-three models use the tag-default context; `llama3.1:70b` and `mixtral:8x22b` use an explicit 4,096-token server context to fit the KV cache. These batteries ran on NVIDIA A100 GPUs; the separate intervention study's RTX PRO 6000 configuration is reported in Section S1.11. No closed-source API is used. A full observational evaluation takes approximately 12 hours.

Corrections and replay checks. Episode-wide source uniqueness was enforced after 20 ambiguous probes were identified in 11 source-monitoring episodes. Those episodes were regenerated and re-inferred for all 55 models (605 evaluations); the other 39 episodes were rescored from stored responses, and all 2,750 final scores were replayed under the current scorer. The image scorer and renderer were corrected for blank-response credit, unanchored location mentions, and font fallback. The final frozen seed-42 image set has matched label distributions for 100 Stroop trials, an exact target-direction by congruency factorial for 100 Flanker trials, and 50 false-belief stories rendered as single four-panel montages. A paradigm-aware parser accepts exact labels or uniquely anchored answers; blank, ambiguous, and unanchored responses score as incorrect. All six VLMs were re-evaluated on this set. A paired replay of all 1,500 items on an RTX PRO 6000 node reproduced 97.9% of item-level scores (1,468/1,500) and preserved the qualitative conclusions.

Grouping	MMLU	ARC-C	GSM8K
WM	0.60*	0.49	0.40
Control	**0.78***	0.59*	**0.60***
Episodic	0.63*	**0.76***	0.61*
ToM	0.74*	0.50	0.51
Meta	0.67*	0.56*	0.42

Table S3: Bivariate Spearman ρ between CogArena grouping scores and external benchmarks. * = $p < 0.05$ after Benjamini-Hochberg correction across the 15 cells (10 of 15 significant).

External-score regime. Table S3 uses official full-precision external-benchmark scores from vendor reports, blogs, and `bf16` model cards, so its correlations mix serving regimes.

S1.4 Per-Paradigm Scaling

Figure S2 summarizes the Pearson correlations with model size, and Figure S3 shows the underlying per-model data for each paradigm across the 20 text LLMs.

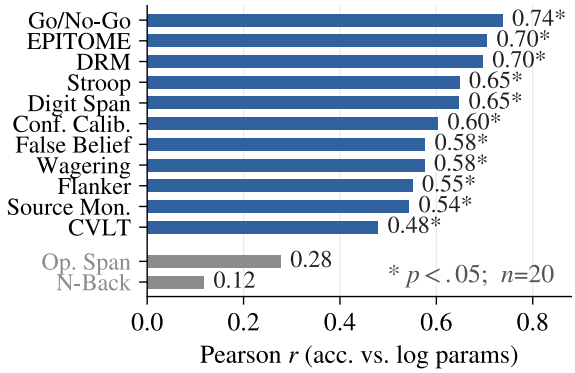

Figure S2: Scaling correlation between log parameter count and accuracy for the 13 paradigms across the 20 text LLMs. Response inhibition, episodic recognition, and theory of mind scale strongly. N-back scales weakly, while CVLT scales moderately under recall-based scoring. Gray bars mark the two nonsignificant correlations, operation span and n-back. Ordering is preserved in the 55-model pool.

Family-random-intercept scaling fits. The models below use maximum likelihood with accuracy regressed on $\log_{10}$ parameter count and a random intercept for model family (20 checkpoints, 11 families, seven represented by one checkpoint). Slopes are therefore accuracy-unit changes for a tenfold parameter increase. Wald statistics are two-sided; the two non-converged fits are retained only as diagnostics.

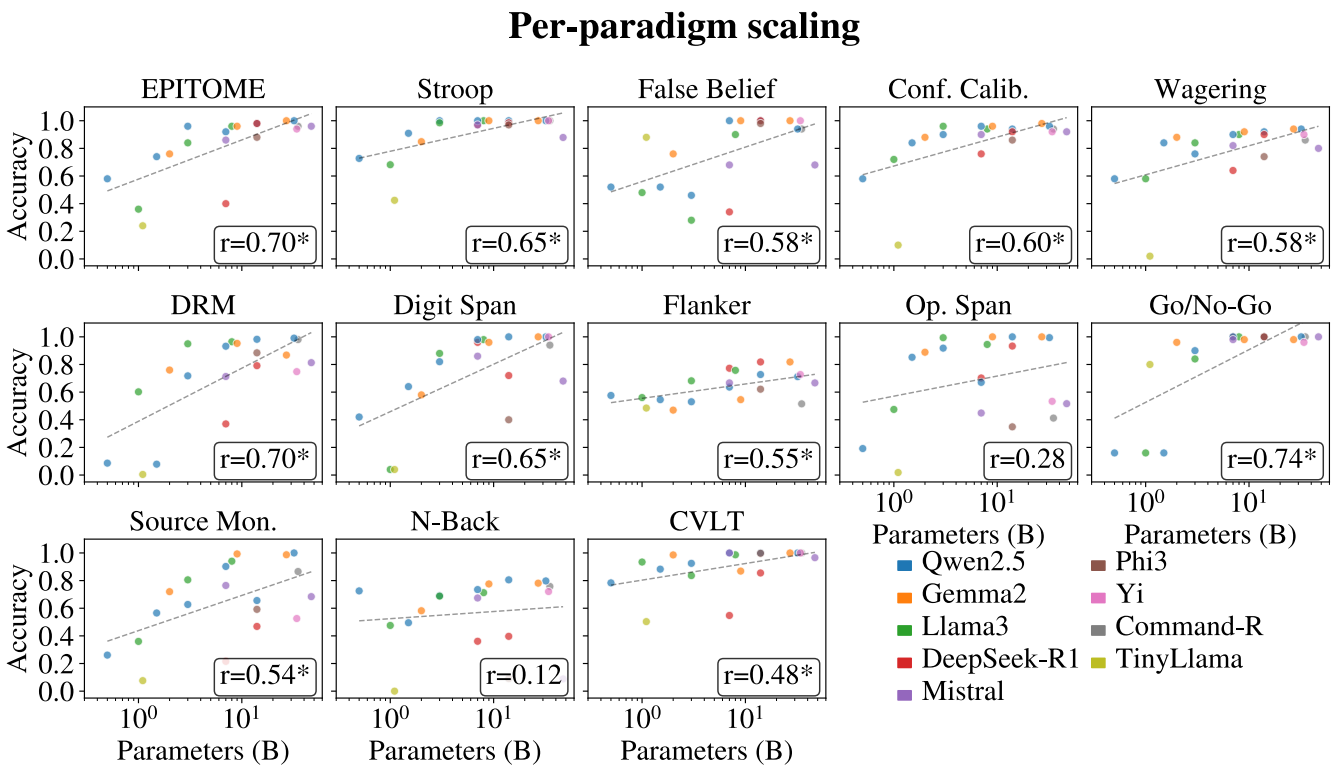

Figure S3: Per-paradigm scaling across the 20 text LLMs. Points show checkpoints, colors show model families, dashed lines are fitted log-size trends, and each panel reports Pearson r. Detailed estimates appear in the adjacent tables.

Paradigm	$\hat{\beta}$	SE	p	V_f	V_e	ICC	Conv.	Bnd.
N-back	.127	.047	.006	.0735	.0069	.914	Yes	No
Digit span	.361	.089	<.001	.0131	.0404	.245	Yes	No
Operation span	.250	.105	.017	.0579	.0368	.611	Yes	No
Stroop	.169	.039	<.001	.0081	.0062	.567	Yes	Yes
Flanker	.114	.039	.003	.0028	.0057	.325	Yes	Yes
Go/No-Go	.381	.083	<.001	.0004	.0388	.011	Yes	Yes
CVLT	.124	.044	.005	.0083	.0083	.501	Yes	Yes
DRM	.384	.077	<.001	3.49×10^{-5}	.0501	.001	No†	Yes
Source monitoring	.327	.067	<.001	.0400	.0156	.720	Yes	No
False belief	.249	.079	.002	2.02×10^{-8}	.0399	.000	No†	Yes
EPITOME	.293	.062	<.001	.0046	.0213	.177	Yes	Yes
Calibration	.201	.042	<.001	.0276	.0060	.822	Yes	No
Wagering	.199	.048	<.001	.0290	.0078	.788	Yes	No

†The DRM and false-belief optimizers did not converge, so their coefficients and Wald values are diagnostic. "Boundary" records the optimizer's boundary warning; complete warnings, log likelihoods, software versions, and machine-readable coefficients are in the code/data supplement.

S1.5 Restricted-Range Robustness

A positive manifold and the within-minus-cross result could in principle be artifacts of restricted-range paradigms because columns with little spread (floor or ceiling) attenuate and distort correlations. We test this directly (Table S4). Empirically, the lowest-variance paradigms are Stroop (SD 0.13), Flanker (0.13), and confidence calibration (0.14), all ceiling-bound; the weakly scaling n-back, the moderately scaling CVLT, and Go/No-Go instead carry substantial variance (SD 0.26, 0.28, 0.27; broad score ranges), so they are not floor- or ceiling-restricted. Across six conditions (dropping the three lowest-variance paradigms; dropping n-back, CVLT, and Go/No-Go; and leaving each of CVLT, Go/No-Go, and n-back out individually), the first principal component remains dominant (50–58% of variance), the raw within-minus-cross gap is positive throughout ($\delta = 0.05$–0.10) and reaches nominal one-sided significance in some conditions (p as low as .04), and the residualized contrast is nominally significant in every condition (δ up to 0.24, one-sided $p \leq .04$). The positive manifold is therefore not a product of restricted range; if anything it strengthens once the high-variance multi-turn paradigms are removed. A separate joint exclusion removes the three paradigms with the most consequential interpretation caveats, namely text Stroop, Go/No-Go, and CVLT. In that 10-paradigm matrix, accuracy separation strengthens to $\delta = .147$ (two-sided $p = .021$; six within-grouping and 39 cross-grouping pairs), whereas construct-native separation remains inconclusive at $\delta = .095$ ($p = .441$). No family-clustered interval was computed for this joint deletion, so it is a countervailing post-hoc sensitivity rather than a replacement primary analysis. The dependence on designed difficulty is reported as a separate 11-paradigm sensitivity panel in the main-paper dimensional-structure analysis. Point estimates were positive at each difficulty tier ($\delta = 0.117$, 0.140, and 0.169; nominal one-sided $p = 0.033$, 0.013, and 0.003),

Condition	PC1	raw δ (p)	resid. δ (p)
Full (13 paradigms)	0.50	+0.081 (.037)	0.216 (.004)
Drop NB, CV, GN	0.58	+0.084 (.048)	0.239 (.025)
Drop 3 lowest-var.	0.51	+0.085 (.124)	0.179 (.039)
Leave out CVLT	0.53	+0.074 (.061)	0.236 (.006)
Leave out Go/No-Go	0.52	+0.096 (.038)	0.216 (.017)
Leave out n-back	0.52	+0.047 (.162)	0.188 (.013)

Table S4: Restricted-range robustness (55 models; 5,000-permutation Monte Carlo label test, seed 42; p one-sided, so values near .05 carry Monte Carlo resolution of about .003). NB n-back, CV CVLT, GN Go/No-Go. PC1 = fraction of paradigm-score variance on the first principal component. raw δ = within minus cross mean paradigm correlation; resid. δ = same after residualizing each paradigm on overall competence (row mean), the general-factor removal that avoids the PC1-orthogonality artifact. The positive manifold (0.50–0.58) holds in every condition; the raw gap is positive throughout and reaches nominal one-sided significance in some conditions, and the residualized contrast is nominally significant in every condition.

and merged-family intervals excluded zero but all included the prespecified 0.15 threshold. All conditions use the same label-permutation test as the headline analysis (seed 42), with 5000 permutations per condition.

S1.6 Construct-Native Rescoring

Raw accuracy also captures shared response-format variance, so the positive manifold and within-grouping gap may combine construct and method effects. We therefore rebuilt the paradigm matrix from the stored responses of the same runs, replacing accuracy with a construct-native score for the seven paradigms that admit one (Table S5); the other six paradigms keep their accuracy scores, and every metric is oriented so that higher means more of the intended ability. For n-back, per-turn responses are re-coded under the strict scorer's parsing rules and unparseable turns are dropped; five small models retain too few parseable turns for a defined d' and are mean-imputed (TinyLlama-1.1B, Phi3-3.8B, Qwen3-4B, StableLM2-1.6B, Starling-7B).

The boundary conclusion remains under construct-native scoring (Table S6). The within-minus-cross gap moves from $+0.08$ (accuracy) to -0.02 (two-sided p=.76; -0.03, p=.67 under canonical scoring), the construct-side threshold sweep certifies equivalence at any margin above 0.051 (0.066 with raw family labels), far below the pre-specified 0.15, and leave-one-family-out gaps stay negative throughout (δ within $[-0.06, -0.01]$, smallest one-sided p=.50). The first principal component's share falls from 50% to 40%, as expected once a shared answering-ability component is removed, while row-mean residualization of the z-scored construct matrix remains null at $\delta = +0.03$ (two-sided p=.68), and removing the first principal component gives $\delta = +0.13$ (p=.037; p=.056 under canonical scoring), a statistic whose null false-positive rate is near-nominal in our simulations (.054/.061 under general-factor-only worlds, both CIs covering .05); because it crosses significance between scoring specifications, we treat it as suggestive rather than as evidence of separable profiles. The null is also not an artifact of unreliable difference scores. Split-half reliabilities of the construct scores (Table S5) are lower than those of accuracies computed on the same items, as expected for difference and signal-detection scores, but remain well above interpretability floors. The construct scores do reorder models, most sharply for DRM, where the construct score correlates negatively with the paradigm's accuracy across models ($r = -0.40$; n-back 0.18, Flanker 0.21, the rest 0.45–0.87). An accuracy profile and a construct profile can therefore disagree paradigm by paradigm, which reinforces the practice conclusion of the main text, while neither establishes a stable, scoring-invariant grouping structure.

S1.7 Simulation-Based Validation of the Separability Test

Three generative worlds calibrated to the accuracy matrix (per-paradigm general-factor loadings fit by least squares to the observed correlations; 55 simulated models per repetition; the same raw, row-mean-residual, and PC1-removal pipeline with label-permutation tests, seed 42) benchmark what the analysis reports when the truth is known. In a pure general-factor world (1,000 repetitions) row-mean residualization has type-I rates .026–.031 at nominal .05, while PC1 removal is near nominal at .054/.061 and the residual signals observed in the real data are $\leq$0.3% tail events, so they are not orthogonalization artifacts. A world adding a text-method factor to the general factor is observationally identical to the pure general-factor world by construction, confirming that a single modality cannot separate the two. Worlds adding five group factors of increasing strength give the power curve in Table S7. A within-grouping correlation increment of 0.15, the profiling threshold, is detected by the raw test with 92% power (the realized raw δ at that simulated arm averages .11), and the residual contrasts observed in the real data correspond to an increment of roughly 0.05–0.10. Horn parallel analysis retains one component for the accuracy matrix and two for the construct-scored matrix, but the second construct component separates the difference- and d'-scored paradigms from the accuracy-scored ones across grouping boundaries. Its strongest loadings are n-back -0.51 and Flanker -0.50 versus CVLT $+0.33$ and source monitoring $+0.24$, indicating a metric-type method factor rather than cognitive structure. The raw gap is also robust to family structure. Equal-family weighting over the 24 merged families gives δ=0.08 (one-sided p=.099, two-sided p=.18), and leaving out any single family keeps δ within [0.06, 0.09] (smallest one-sided p=.017). Framed as a threshold sweep, the family-clustered interval certifies equivalence only at margins above 0.145 for the accuracy matrix (0.174 with raw family labels), so the pre-specified 0.15 margin is met under merged labels but not under raw labels; for the construct matrix equivalence holds at any margin above 0.051 (0.066 raw). A joint family$\times$item bootstrap resamples the 24 merged families and, within each replicate, every paradigm's items (20,000 replicates per seed, seeds 42–44). The two-level 95% CI for δ is $[-0.015, 0.151]$ under seed 42, $[-0.017, 0.152]$ under seed 43, and $[-0.014, 0.150]$ under seed 44; family-

Paradigm	Construct score	$SB_{constr.}$	$SB_{acc.}$
Stroop	interference, $acc_{incong.} - acc_{cong.}$	.70	.95
Flanker	interference, $acc_{incong.} - acc_{cong.}$	.65	.81
Go/No-Go	d' (log-linear)	.97	.99
n-back	d' over match/no-match turns	.99	1.00
DRM	$-(FA_{critical\ lure} - FA_{unrelated})$	.98	Not defined
Conf. calibration	1−Brier (fixed-scorer correctness)	.79	.90
Wagering	type-2 d' (wager and correctness coupling)	.67	.91

Table S5: Construct-native scores and split-half reliabilities (Spearman-Brown, median over 100 random splits). $SB_{acc.}$ uses accuracy on the same split units. The other six paradigms retain accuracy; the DRM accuracy baseline is undefined on these units.

	Accuracy scores	Construct scores
within − cross δ	+0.08	−0.02
permutation p (one/two-sided)	.036/.057	.60/.76
family-clustered 95% CI	[−0.012, 0.145]	[−0.11, 0.05]
PC1 share	50%	40%
z-scored row-mean residual δ (p)	0.22 (<.01)	0.03 (.68)
PC1-removal sensitivity δ (p)	0.27 (<.01)	0.13 (.04)

Table S6: Separability under accuracy and construct-native scoring for 55 models. Both columns use the same 50,000-permutation test and 5,000-resample family bootstrap. The main-text two-level and canonical intervals do not exclude $\delta = .15$, while the construct interval does. Row-mean residualization and PC1 removal are sensitivities.

w	incr.	raw δ	P(raw sig.)	res. δ	P(PC1 sig.)
0 (pure g)	0	−.04	.00	−.01	.00
0.15	.02	−.02	.00	.04	.00
0.22	.05	.01	.01	.10	.00
0.32	.10	.06	.42	.23	.41
0.39	.15	.11	.92	.35	.96
0.45	.20	.16	1.00	.46	1.00

Table S7: Generative benchmarks for the separability pipeline (500–1,000 repetitions per row). incr. is w^2; raw and res. δ are mean raw and residual gaps. The P columns give raw-test detection and observed-PC1-pattern replication rates. At $w^2 = .15$, detection is 92% while realized δ averages .11.

only resampling gives $[-0.012, 0.145]$ and raw family labels $[-0.026, 0.174]$ (seed 42). The two-level upper limits sit at the 0.15 margin, so equivalence at the pre-specified threshold is not robustly excluded once both variance sources are resampled jointly.

S1.8 Cross-System Comparison

The cross-modal stress test, run on the three paradigms with a visual form, shows that text adaptation can alter a paradigm's construct (Figure S4). Image Stroop shows a human-direction accuracy congruency effect absent in the text version (MacLeod 1991); for example, Qwen2.5-VL scores 100%/84% and LLaVA-7B 100%/0% on congruent/incongruent trials. Image false belief remains heterogeneous across the six VLMs, from 66% for Qwen2.5-VL to 0% for LLaVA-7B and Moondream, under a strict parser that accepts exact or uniquely anchored answers and scores blank or unanchored responses as incorrect.

S1.9 Human Comparison

Matched human data exist for only two paradigms. Strachan et al. (2024) report near-ceiling (>95%) adult 1st-order false belief on text stories (vs. our 85.2%/68.4% for 1st/2nd order on procedurally generated scenarios), and Jones, Trott, and Bergen (2024) report adult EPITOME performance well above chance. For the other 11 paradigms (especially text-adapted executive-function tasks) no matched human data on text-based versions exist, a field-wide gap.

Quantitative battery-wide comparison between humans and LLMs is not warranted for the other 11 paradigms, whose source studies report reaction time, span length, or metacognitive measures rather than comparable accuracies. We therefore use human studies only to define directional signatures (e.g., congruent versus incongruent, or easier versus harder load) and do not infer a cross-species level or difficulty-profile ranking. Under the corrected scorer, model operation-span accuracy averages 69.2%; this number is not commensurate with human span length.

S1.10 Contamination Analysis

We test contamination for 5 models (Qwen2.5 0.5B/7B/32B, Gemma2-9B, DeepSeek-R1-14B) × 10 single-turn paradigms using Fisher's exact test on correct/incorrect counts, comparing $n = 30$ classic items per paradigm (canonical or widely reproduced stimuli likely to appear in training corpora, e.g. the original Sally-Anne scenario) against $n = 30$ procedurally generated novel items. Of 50 combinations, only Qwen2.5-0.5B on Stroop reaches uncorrected

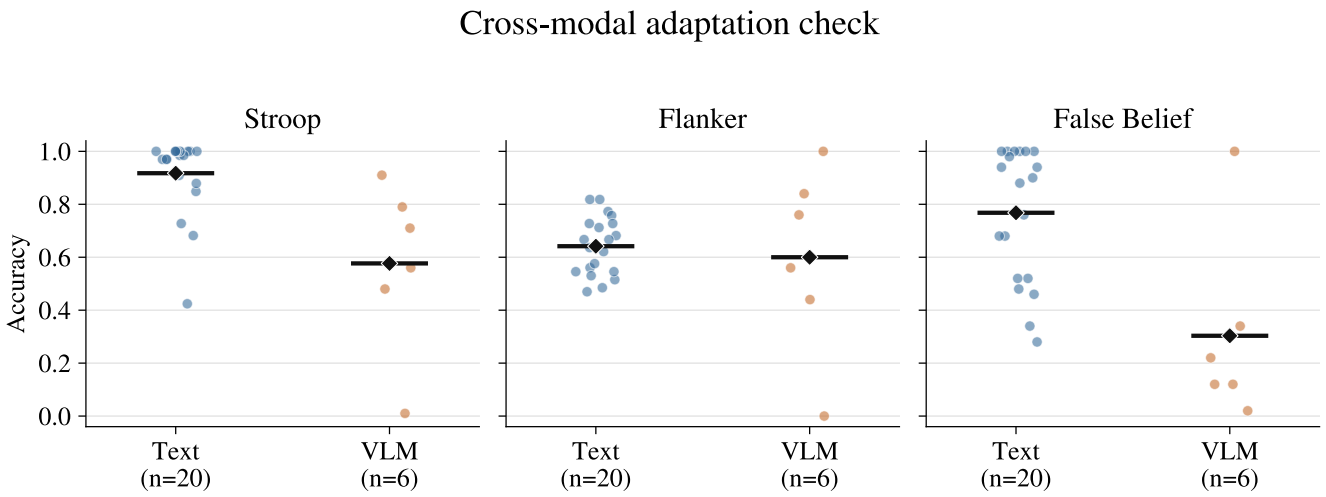

Figure S4: Unpaired descriptive score distributions for 20 text LLMs and six VLMs on three shared paradigms. Dots are checkpoints and black diamonds with horizontal lines mark pool means. The pools differ, so between-pool gaps are not paired modality effects. Blank VLM completions count as incorrect.

significance (classic 100% vs. novel 77%, $p = 0.011$), which does not survive Bonferroni correction ($\alpha_{\text{adj}} = 0.001$). The remaining four models show zero flagged paradigms. Note that the result JSON files also flag paradigms with gap $>$ 10% as `contamination_detected` regardless of statistical significance; only the Fisher test p-values should be used for inference. The probe therefore detects no correction-surviving classic-item advantage, but it is not powered to exclude small contamination effects. We use only procedurally generated items in main evaluation.

S1.11 Fully Crossed Intervention and Family Prediction

Design and scope. This study was designed after the observational analysis and is not a preregistration of the original benchmark result. Before formal intervention outcomes were inspected, we froze the model panel, held-out item manifest, prompts, estimands, seeds, thresholds, and all-nine decision rule. The panel contains two checkpoints from each of six families. They are Qwen2.5 (3B, 14B), Gemma2 (2B, 9B), Llama2 (7B, 13B), Gemma3 (12B, 27B), Falcon3 (7B, 10B), and OLMo2 (7B, 13B). Each checkpoint receives 18 new items per paradigm (six per difficulty) under seven conditions consisting of baseline, a length-matched neutral placebo, and five answer-free scaffolds targeting working memory, cognitive control, episodic source binding, agent belief states, or metacognitive forecasting. Every scaffold is applied to every paradigm, yielding $12 \times 13 \times 18 \times 7 = 19{,}656$ model-item-condition evaluation records. Exact scaffold text is included in `PREPILOT_SPEC.json`; none contains an item answer.

All models were served on the c04 RTX PRO 6000 node at a 4,096-token context with greedy decoding, a 512-token completion ceiling, and `reasoning_effort=none`. The same held-out item is paired across conditions. Primary scoring uses strict-v4 operation-span recall, unique-hit capped CVLT recall, strict n-back turn scoring, and the frozen native scorer elsewhere; canonical whitespace operation-span scoring is a sensitivity. Protocol-invalid completions are retained under intention-to-treat scoring as zero. The formal run completed 19,656 records with a maximum condition-level protocol-invalid rate of .00392.

Estimand and inference. For targeted intervention j, let its item-mean accuracy gain over the neutral placebo be G_{jp}. We define

$$S_j = \text{mean}_{p \in \mathcal{M}_j} G_{jp} - \text{mean}_{p \notin \mathcal{M}_j} G_{jp},$$

$$\Gamma = \frac{1}{5} \sum_{j=1}^{5} S_j.$$

where $\mathcal{M}_j$ is the frozen set of paradigms matched to intervention j. Models, paradigms, and interventions receive equal weight. The primary interval uses 20,000 crossed bootstrap draws that resample six families with replacement while retaining both checkpoints and independently resample 18 items within each paradigm using the same draw across conditions. An exact correspondence test enumerates all $5! = 120$ intervention-to-group mappings. A six-fold family-LOFO ridge-logistic comparison asks whether five diagonal terms improve held-out soft-Bernoulli log likelihood beyond placebo accuracy and additive intervention, paradigm, and difficulty terms.

Primary and family results. The aggregate diagonal tendency is $\Gamma = .0199$ (Table S8); canonical operation-span scoring gives .0198 with the same exact $p = .0167$. Family estimates are Qwen2.5 .0176, Gemma2 $-.0007$, Llama2 .0447, Gemma3 .0209, Falcon3 .0362, and OLMo2 .0007. Thus five of six are positive, but an exact sign test is coarse (one-sided $p = .109$), and an exact family sign-flip test gives one-sided $p = .031$ and two-sided $p = .063$.

Scaffold	Matched grouping	S_j	95% CI	Mapping p_{BH}
Ordered ledger	Working memory	.0448	[.0055,.0899]	.50
Rule rehearsal	Cognitive control	.0038	[−.0414,.0493]	.60
Source binding	Episodic memory	.0309	[−.0124,.0860]	.50
Belief-state ledger	Theory of mind	.0257	[−.0116,.0698]	.60
Forecast and calibrate	Metacognition	−.0056	[−.0487,.0323]	.60
Equal-intervention mean Γ		.0199	[.0041,.0360]	exact $p = .0167$

Table S8: Intervention selectivity relative to the length-matched neutral placebo. Individual mapping p values have only five distinct assignments and are BH-adjusted; the omnibus exact mapping test for Γ enumerates all 120 mappings. Intervals are crossed family-by-item percentile intervals.

Family-LOFO Γ remains .0150–.0240, yet adding the diagonal terms does not improve predictive likelihood. Total $\Delta LL = -.904$, and only Qwen2.5 and Gemma3 improve. Across the seven conditions, PC1 continues to explain 57.6–66.8% of paradigm variance. Medium-difficulty items show the clearest exploratory tendency; the hard-minus-easy contrast is approximately zero.

Alternate-wording replication. After the frozen study, we repeated the same crossed design with an alternate wording for each targeted scaffold. The main text compares the two wordings in a compact table. The replication retained the same 12 checkpoints, six families, held-out items, scoring rules, estimand, and gates. It produced a smaller positive diagonal estimate ($\Gamma = .0134$, crossed 95% CI [−.0030,.0298]; exact mapping $p_2 = .0333$), with four of six family estimates positive. Selective terms improved aggregate family-LOFO likelihood by +.771, but only three of six held-out families improved. Six of nine gates passed; the crossed-interval, empty-response, and operation-span parse exclusions failed. The all-gates decision therefore remains FAIL.

Post-hoc audit diagnostics. Six post-hoc analyses characterize robustness without entering the frozen decision. First, all 13 leave-one-paradigm-out values remain positive (.0164–.0237). Second, a hierarchical bootstrap over families, paradigms within groups, and items gives a 95% interval of [.0007,.0420]; the 2–3 observed paradigms per theoretical stratum make this a design sensitivity rather than strong paradigm-population inference. Third, observable evaluability (protocol-valid, nonempty, and parseable for operation span) has $\Gamma = .0072$, CI [−.0067,.0246]. A linear accounting identity attributes .0130 of the .0199 accuracy contrast to pairs where both sides are evaluable and .0062 to target-only-evaluable pairs; the remaining .0008 is the net of placebo-only (−.00006) and neither-evaluable (+.00086) contributions. This decomposition is descriptive rather than mediational. Fourth, neutral placebo versus baseline is −1.29 accuracy points, CI [−3.14,.25], and −1.60 evaluability points, CI [−3.31,−.07]. Fifth, replacing placebo with the no-scaffold baseline gives $\Gamma = .02067$ (crossed family-by-item CI [.00342,.03815], exact 5! mapping $p = .0167$). The group-differential placebo contribution is −.00075 (CI [−.00375, .00241]), and targeted arms average .81 accuracy points below baseline. Together, these comparisons preserve the diagonal tendency across reference conditions while distinguishing it from an overall accuracy benefit. Sixth, the exact six-family tests and every family estimate are reported above. The frozen decision remains FAIL.

Freeze and reporting amendment. The intervention protocol is outcome-frozen rather than a public preregistration. Before aggregate results were released, an outcome-blind reporting amendment defined how the frozen rule handles unestimable minimum-cell sensitivities. It changed only the reporting status of those gates. The code archive contains the frozen specification, amendment, run manifest, analyzers, aggregate outputs, and SHA-256 manifests; raw response text is omitted from the anonymous repository.

S1.12 Post-hoc Profile Transport and Stability

Three diagnostics, all post-hoc and outside the frozen intervention rule, test alternative explanations for the boundary result. First, after centering checkpoints within each of the 11 families represented by multiple models, the accuracy-based grouping contrast is nearly zero ($\delta = .0105$, exact two-sided $p = .798$; family-bootstrap 95% CI [−.087, .085]). Across 24 held-out family centroids, adding the other paradigms from a target's proposed grouping to a general-component predictor does not reduce prediction error (relative RMSE gain −1.76%, family-bootstrap CI [−6.30%, 2.01%]; 3/13 target paradigms improve). Under construct-native scores, the corresponding gain is −4.73% (CI [−6.13%, −2.87%]; 0/13 improve). These results concern this finite model-family panel and are not population estimates over future architectures.

Second, replay stability is high. For 20 models and eight eligible single-turn paradigms, 8,420 same-item response pairs from adjacent greedy-decoding administrations give absolute-agreement ICC(A,1)=.979 for model-centered profile cells (family-bootstrap CI [.961,.990]); the mean within-model profile correlation is .981 (CI [.962,.992]). This diagnostic measures identical-item response and serving stability; construct validity is evaluated separately by the structural and transport analyses. Third, family-centroid structure remains descriptive. Across all 24 families, $\delta = .079$ (exact two-sided $p = .184$), whereas construct-native family centroids give $\delta = -.091$. The released scripts and manifests bind every matrix, resample count, seed, and eligibility exclusion used here.

Gate	Result	Evidence
Crossed-bootstrap lower endpoint > 0	PASS	CI [.0041,.0360]
At least four of six family estimates > 0	PASS	5/6 positive
Held-out-family selective $\Delta LL > 0$	FAIL	$-.904$; 2/6 folds improve
Exact one-sided mapping $p \leq .05$	PASS	$p = .0167$
Every condition's protocol-invalid rate $\leq 1\%$	PASS	maximum .392%
Protocol-invalid paired exclusion preserves $\geq .5\Gamma$ and three items per cell	PASS	$\Gamma = .01965$, ratio .986, minimum 14
Empty-response paired exclusion preserves $\geq .5\Gamma$ and three items per cell	FAIL	855 pairs excluded; minimum cell 0; unestimable
OSpan parse-none paired exclusion preserves $\geq .5\Gamma$ and three items per cell	FAIL	262 pairs excluded; minimum cell 0; unestimable
Response-length adjustment preserves $\geq .5\Gamma$	PASS	$\Gamma = .01847$, ratio .927

Table S9: Frozen all-required confirmation rule. The decision is FAIL because three of nine gates fail. Unestimable exclusions are scored FAIL under the frozen minimum-cell rule.

S2 Gymnasium Environment API

Every paradigm is exposed as a registered `gymnasium.Env` (Gymnasium 1.x) with text observation and action spaces (`spaces.Text`) and the standard five-tuple `step` returning `(observation, reward, terminated, truncated, info)`. An environment is created with `gym.make` (Table S10) and driven by the usual `reset(seed)`/`step(action)` loop; the reward is the per-turn partial match of the response against the expected answer, and `env.score()` returns episode accuracy. Single-turn paradigms are one-step episodes, while the multi-turn working- and episodic-memory paradigms (n-back, operation span, CVLT) run their full turn sequence. The environments reuse the same procedural item generators as the batch evaluation, so an agent driven through the Gymnasium loop sees identical items.

S3 Pilot Agent Evaluation

In a pilot agent evaluation with 4 models (Qwen2.5-7B/32B, DeepSeek-R1-14B, TinyLlama-1.1B), agents with tool access achieve 100% on false belief (all 4). N-back performance is 100% for Qwen2.5-7B/32B and 50% for the other models. WCST remains challenging (25% mean). The same model may produce different per-paradigm patterns depending on the evaluation interface. TinyLlama scores 88% on text false belief but 100% in agent mode, suggesting external memory tools partially compensate for parametric limitations. Agent-mode answers are graded by whether the expected answer appears in the final response, a lenient criterion that can credit restated options, so these pilot numbers are upper bounds. Larger-scale agent evaluation is needed to confirm these patterns.

S4 Positioning Relative to Prior Evaluations

Table S12 tabulates how CogArena relates to the closest LLM cognitive, psychometric, and scaffold-validity evaluations discussed in Section 2.

Grouping	Environment id	Turns
Working Memory	`CogArena/DigitSpan-v0`	S
	`CogArena/NBack-v0`	M
	`CogArena/OperationSpan-v0`	M
Cog. Control	`CogArena/Stroop-v0`	S
	`CogArena/Flanker-v0`	S
	`CogArena/GoNoGo-v0`	S
Episodic Mem.	`CogArena/DRM-v0`	S
	`CogArena/SourceMonitoring-v0`	S
	`CogArena/CVLT-v0`	M
Theory of Mind	`CogArena/FalseBelief-v0`	S
	`CogArena/EPITOME-v0`	S
Metacognition	`CogArena/ConfidenceCalibration-v0`	S
	`CogArena/Wagering-v0`	S

Table S10: The thirteen registered Gymnasium environments, one per paradigm. The turns column uses M for a multi-turn episode and S for a single-turn episode.

S5 Paradigm Inventory and Per-Paradigm Results

Table S13 lists the 13 paradigms with their groupings, source literature, human sample sizes, adaptation ratings, and evaluation modes.

Table S14 gives per-paradigm accuracy for representative models on the 10 single-turn paradigms.

Paradigm	Adapt.	Matched human	Per-model signature	Scaling	Scorer-sens.	Recommended use
Digit Span	Low	No	aggregate	strong (r=0.65)	No	WM capacity probe
N-Back	Low	No	load effect (15/20)	near-zero (r=0.12)	Yes	WM updating; strict-scorer caveat
Operation Span	Low	No	aggregate	weak (r=0.28)	Yes	dual-task WM probe
Stroop (text)	Med	No	weak (7/20)	strong (r=0.65)	No	prefer image version
Flanker	Med	No	replicates (18/20)	moderate (r=0.55)	No	reliable conflict probe
Go/No-Go	Med	No	aggregate (84% GO)	strong (r=0.74)	No	rule accuracy; base-rate caveat
DRM	Low	No	false memory replicates (18/20)	strong (r=0.70)	No	false-memory probe
Source Monitoring	Low	No	graded difficulty (aggregate)	moderate (r=0.54)	No	source-attribution probe
CVLT	Low	No	aggregate	moderate (r=0.48)	Yes	visible list; availability caveat
False Belief	Low	Yes	weak (12/20, n.s.)	moderate (r=0.58)	No	ToM; matched human data exists
EPITOME	Low	Yes	replicates (25/35, expansion pool)	strong (r=0.70)	No	ToM; matched human data exists
Conf. Calibration	Low	No	aggregate	moderate (r=0.60)	No	metacognitive monitoring
Wagering	Med	No	aggregate	moderate (r=0.57)	No	metacognitive control

Per-model signature replication is tested for the 6 condition-split paradigms (N-Back, Stroop, Flanker, False Belief, EPITOME, DRM); all other paradigms support an aggregate-level behavioral signature only.

Table S11: Per-paradigm validity ledger. Adapt. is adaptation distance. Matched human denotes comparable text-version accuracy. Signature counts report checkpoint-level directional replication; Section S1.2 gives the family sensitivity. Scaling is Pearson r between $\log$(parameters) and accuracy. Scorer-sens. marks material changes under an alternative scorer.

Work	Cog.-sci. tasks	Proc. gen.	Construct checks	Conv. and discr. validity	Adapt. map	Crossed specificity	Family-heldout
CogBench (Coda-Forno et al. 2024)	●						
Centaur (Binz et al. 2025)	●						
Momentè (Momentè et al. 2025)	●		○				
NeuroCognition (Haznitrama, Ardi, and Oh 2026)	●	○					
Jung et al. (2026)	●		●	○			
de Langis et al. (2026)	●			○			
Ilić and Gignac (2024)				○			
Burnell et al. (2023)				○			
ADeLe (Zhou et al. 2026)			○				
Serapio-García et al. (2025)			●	○			
ActTraitBench (Yang et al. 2026)	●		●	○	○		
Contreras (2026)			●	○			
Bugaud (2026)	●	●					
Trott, Rivière, and Jones (2026)	○		○	○			
NeuReasoner (Javadov et al. 2026)	●		○			○	
Beyond Direct Gains (He et al. 2026)			○			○	
CogArena (ours)	●	●	●	●	●	●	●

Table S12: Comparison with the closest evaluation lines. ● = yes, ○ = partial, blank = not reported. "Crossed specificity" requires a theory-matched intervention to be tested on both matched and off-target proposed groupings against a neutral control; "model-family-heldout" requires profile or selective terms to predict checkpoints from unseen model families. In this comparison, CogArena alone combines repeated cognitive paradigms, convergent and discriminant validity, fully crossed intervention specificity, and held-out-model-family prediction for the same theory-motivated taxonomy.

Grouping	Paradigm	Human Norm Source	N_{human}	Adapt.	Modes
Working Memory	Digit Span	WAIS-IV (Wechsler 2008)	2.2K	Low	T
	N-Back	(Pelegrina et al. 2015) (verbal 1–3-back child norms)	3,722‡	Low	T
	Operation Span	(Redick et al. 2012) (automated complex spans)	6K+	Low	T
Cog. Control	Stroop	(Van der Elst et al. 2006) (adult norms)	1,856	Med	T, V
	Flanker	(Eriksen and Eriksen 1974)	12	Med	T, V
	Go/No-Go	(Votruba and Langenecker 2013)	276	Med	T
Episodic Mem.	DRM False Memory	(Roediger and McDermott 1995)	66	Low	T
	Source Monitoring	(Johnson, Hashtroudi, and Lindsay 1993) (review)	Not reported	Low	T
	CVLT Word List	(Delis et al. 2000) (CVLT-II)	1,087	Low	T
Theory of Mind	False Belief	(Wellman, Cross, and Watson 2001); (Strachan et al. 2024) (single-task analysis)	49†	Low	T, V, A
	EPITOME	(Jones, Trott, and Bergen 2024) (six component studies)	44–1,156	Low	T
Metacognition	Conf. Calibration	(Fischhoff, Slovic, and Lichtenstein 1977) (five studies)	528	Low	T
	Wagering	(Persaud, McLeod, and Cowey 2007) (three experiments)	67	Med	T

Entries marked "review" aggregate many studies without a single sample. Classic within-subject paradigms such as Flanker attain statistical power through many trials per participant rather than large samples. EPITOME recruitment varied across its six component studies; the range shown is the smallest and largest recruited sample. The wagering total combines 66 students across two experiments and one blindsight participant.
† (Strachan et al. 2024) report N=49 for the original-versus-novel false-belief analysis; N=1,907 is the total across their full ToM battery.
‡ (Pelegrina et al. 2015) enrolled 3,722 children aged 7–13; 3,296 completed 2-back and 2,141 completed 3-back under the study's performance-contingent progression rule.

Table S13: CogArena paradigm inventory. Each paradigm is adapted from a validated human experiment. N_{human} denotes the participant count, range, or documented total in the cited source. Adapt. denotes adaptation distance, where Low means the core construct is preserved and Med means the mechanism is partially altered. T denotes Text, V denotes VLM, and A denotes Agent.

Model	Size	DS	ST	FL	GN	DRM
TinyLlama	1.1B	4	42	48	80	0
Qwen2.5	0.5B	42	73	58	16	9
Llama3.2	1B	4	68	56	16	60
Gemma2	2B	58	85	47	96	76
Qwen2.5	7B	98	**100**	64	**100**	93
Mistral	7B	86	97	67	98	71
DeepSeek-R1	7B	96	97	77	**100**	37
Llama3.1	8B	98	**100**	76	**100**	96
Gemma2	9B	96	**100**	55	98	95
Qwen2.5	14B	**100**	**100**	73	**100**	98
DeepSeek-R1	14B	72	98	**82**	**100**	79
Gemma2	27B	**100**	**100**	**82**	98	87
Qwen2.5	32B	**100**	**100**	71	**100**	**99**
Yi	34B	**100**	**100**	73	96	75
Command-R	35B	94	**100**	52	**100**	98

Model	Size	SM	FB	EP	CC	WG
TinyLlama	1.1B	8	88	24	10	2
Qwen2.5	0.5B	26	52	58	58	58
Llama3.2	1B	36	48	36	72	58
Gemma2	2B	72	76	76	88	88
Qwen2.5	7B	90	**100**	92	96	90
Mistral	7B	76	68	86	90	82
DeepSeek-R1	7B	22	34	40	76	64
Llama3.1	8B	94	90	96	94	90
Gemma2	9B	99	**100**	96	96	92
Qwen2.5	14B	66	**100**	98	94	92
DeepSeek-R1	14B	47	**100**	98	92	90
Gemma2	27B	99	**100**	**100**	**98**	**94**
Qwen2.5	32B	**100**	94	**100**	96	**94**
Yi	34B	53	**100**	94	92	90
Command-R	35B	87	94	96	94	86

Table S14: Representative single-response paradigm accuracies (%). Bold marks a column maximum. DS = Digit Span, ST = Stroop, FL = Flanker, GN = Go/No-Go, DRM = DRM false memory, SM = source monitoring, FB = false belief, EP = EPITOME, CC = confidence calibration, and WG = wagering. Across all 20 models, column means are 74, 92, 64, 84, 71, 65, 77, 81, 85, and 78%, respectively.