A Cognitive Battery for Foundation Models: Theory-Grounded Benchmarks for Attention, Learning, Metacognition, Executive Function, and Social Cognition

Zacharie Bugaud [1]

Abstract

We present a cognitive benchmark battery for foundation models: five procedurally generated evaluations totalling 25,390 items across 1,138 sub-task types, each operationalising a widely studied family of cognitive constructs: selective attention (Controlled Distractor Injection, CDI), fluid learning (Alien Grammar Induction, ALGIn), metacognitive calibration (Epistemic Calibration Under Uncertainty, ECUU), executive control (Dynamic Rule Override, DRO), and theory of mind (Recursive Belief Tracking, RBT). Rather than report a single accuracy number, each benchmark traces a *degradation profile* along a controlled difficulty axis (distractor count, rule count, recursion order, etc.), turning the evaluation into a parametric probe whose shape can be predicted from a candidate theory of the underlying capability. We describe the design, give worked examples, and discuss how the resulting profiles plug into the workshop's theory–benchmark virtuous cycle. This paper introduces the battery and its rationale; a companion release will report calibration and model scores.

1. Introduction

Most widely used LLM benchmarks emphasise crystallised knowledge (facts and reasoning over familiar patterns (Chollet, 2019)) and report aggregate accuracy on heterogeneous test sets. Aggregate accuracy is opaque: a model that memorises millions of math problems scores well on math benchmarks yet may collapse when problem structure shifts, and the benchmark tells us nothing about *why*.

For the workshop's central question (how to combine theory with benchmarking so that empirical scores are predictable from, and predictive of, structural properties of models), we need evaluations that (i) target a single capability rather than a bundle, (ii) expose a controllable difficulty axis on which a theory can make quantitative predictions (e.g. "accuracy should drop log-linearly in distractor count"), and (iii) are robust to training-set contamination so that the same theory can be tested on the next generation of models.

This paper presents a battery of five benchmarks designed against those three criteria. Each benchmark is anchored in an established experimental paradigm from cognitive psychology. We do not claim that human cognition decomposes cleanly into five core processes, nor that any such taxonomy is settled in the literature; we use psychological paradigms as a source of *well-engineered task templates* (parametric difficulty manipulations that work, documented human baselines, and known confounds to control for), not as a claim about the structure of mind.

The five benchmarks (CDI, ALGIn, ECUU, DRO, RBT) cover selective attention, fluid learning, metacognitive calibration, executive control, and recursive belief tracking. They share a common architecture: procedural generation from typed templates, graded difficulty, and a degradation profile as the primary output. We describe each in Section 2, the cross-cutting design choices in Section 3, and how the resulting profiles plug into the workshop's theory–benchmark loop in Section 4. Section 5 sets out the human baseline strategy and Section 6 states what this paper does and does not claim; in particular, full model evaluation and human calibration are deferred to a companion release.

Scope. We adopt these five construct families as a useful operational partition of the LLM evaluation space, not as a canonical taxonomy of human cognition. Other partitions are defensible; the contribution here is the design and instrumentation of each individual benchmark, and the demonstration that the same template architecture supports five very different capability probes.

[1]Astera Institute, Berkeley, CA, USA. Correspondence to: Zacharie Bugaud <zacharie@astera.org>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.

2. The Five Benchmarks

2.1. CDI: Controlled Distractor Injection (Attention)

CDI targets selective attention by holding a core task constant while parametrically injecting 0–6 irrelevant distractor paragraphs. We vary distractor position (before, after, buried, surrounded) and include an adversarial condition with same-number distractors to defeat lexical heuristics.

Example (simplified, 2 distractors, buried): "*Bob has 12 apples. [A 90-word paragraph about Antarctic penguins.] He gives 4 to Alice. [A 120-word paragraph about supply chains, mentioning the number 7.] How many apples does he have?*" The target is 8; the lure is 7.

CDI draws on filter theory (Broadbent, 1958), the cocktail-party effect (Cherry, 1953), and inattentional blindness (Simons & Chabris, 1999). The position manipulation parallels the "Lost in the Middle" effect (Liu et al., 2024), which gives a concrete predicted shape for the degradation curve.

Stats. 4,938 items, 229 sub-task types, 15 paradigm families, Boolean scoring.

2.2. ALGIn: Alien Grammar Induction (Learning)

ALGIn tests in-context fluid learning via procedurally generated "alien" languages. Each language has a novel phonological inventory (4–8 consonants, 3–5 vowels), a small set of morphological rules (affixation, reduplication, vowel shift, infixation, circumfixation), and a word order drawn from six typologically attested options. The model is shown $k \in \{2, 4, 8, 12\}$ glossed examples and asked to produce or judge new forms.

Example (simplified, plural-suffix rule): "tip = 'cat', tip-na = 'cats'; mor = 'tree', mor-na = 'trees'; kor = 'dog', kor-na = ?" The model should induce the suffix rule and answer *kor-na = 'dogs'*.

Because each language is generated on the fly, training-set overlap is unlikely; the k-step curve gives sample efficiency, paralleling human statistical-learning studies (Saffran et al., 1996) and Cattell's fluid intelligence construct (Cattell, 1963).

Stats. 5,585 items, 223 sub-task types, 14 paradigm families, 10 alien grammars, Boolean scoring.

2.3. ECUU: Epistemic Calibration Under Uncertainty (Metacognition)

ECUU asks models to commit to an answer *and* a confidence in the same turn. Core calibration items are scored with the (inverted) Brier score, a proper scoring rule that jointly rewards accuracy and calibration. Ten difficulty categories range from trivially solvable to provably unknowable; four multi-turn protocols adapt Hart's feeling-of-knowing (Hart, 1965), Fischhoff's hindsight bias (Fischhoff, 1975), the illusion of explanatory depth (Rozenblit & Keil, 2002), and boundary awareness.

Example (provably unknowable item): "*What is the exact number of grains of sand on the beach at Praia da Marinha, Portugal, at 12:00 UTC on 1 January 2025? Give a value and a confidence in [0,1].*" Well-calibrated systems should return very low confidence; overconfident systems are penalised by the Brier score.

The construct follows Flavell's (1979) formulation of metacognitive monitoring as the link between confidence and accuracy.

Stats. 4,975 items, 226 sub-task types, 12 paradigm families, float scoring.

2.4. DRO: Dynamic Rule Override (Executive Function)

DRO adapts neuropsychological executive-function paradigms: perseveration (WCST), inhibitory control (Stroop, Go/No-Go), planning (Tower of Hanoi, grid pathfinding), working memory (N-back, digit span, running memory), and feedback-based learning (Iowa Gambling Task). A core design element is *prepotent completion suppression*, which tests whether a model can override its default next-token continuation.

Example (prepotent suppression): "*Roses are red, violets are ___. Do NOT complete the poem. Instead, output the single word 'green'.*" The correct answer is *green*; the prepotent answer is *blue*. Difficulty scales from 1 to 6 by how strongly the prepotent answer is cued.

The task families map onto Miyake et al.'s (2000) three-factor account of executive function: inhibition, shifting, and updating. The prepotent-suppression family is, to our knowledge, new in the LLM setting and exploits the structural tension between autoregressive generation and instruction following.

Stats. 4,976 items, 230 sub-task types, 12 paradigm families, 6 difficulty levels, Boolean scoring.

2.5. RBT: Recursive Belief Tracking (Social Cognition)

RBT evaluates theory of mind through procedurally generated false-belief scenarios at recursion orders 1–4, with the three classes of anti-heuristic controls identified by Ullman (2023) (true-belief controls, double-move items, indirect access). Additional families cover pragmatic inference, faux-pas detection, emotional ToM, deception detection, and perspective taking.

Example (order 2, false belief): "*Alice puts the keys in the drawer and leaves. Bob moves them to the box without Alice*

seeing. Alice tells Carol she thinks *Bob may have moved the keys. Where does Bob believe that Alice will first look?*" Correct: *the drawer*; common shortcut: *the box*.

RBT extends the Sally–Anne paradigm (Baron-Cohen et al., 1985) to recursive depth and is designed to test the "ToM cliff" hypothesis: that accuracy degrades steeply between orders 2 and 3.

Stats. 4,916 items, 230 sub-task types, 13 paradigm families, recursion orders 1–4, Boolean scoring.

3. Design Principles

Procedural generation. Items are sampled from typed templates, optionally combined with expert-curated content. The same template can produce arbitrarily many items, so the benchmark definition is decoupled from any particular release of items, a property that matters once items leak into training corpora.

Parametric difficulty. Every benchmark exposes a graded difficulty axis (distractor count, rule count, recursion order, prepotent strength, unknowability tier). The headline metric is the degradation profile along that axis, not a single accuracy number. This gives benchmarks the same shape as the dose–response curves they were modelled on.

Anti-heuristic controls. Each benchmark includes items designed to be misclassified by an obvious shortcut: same-number lure distractors in CDI, irregular-form items in ALGIn (which trap rule overgeneralisation), the unknowable tier in ECUU (which penalises surface confidence under a proper scoring rule), prepotent traps in DRO, and double-move items in RBT. Reporting accuracy on the controlled subset alongside the full set surfaces construct-validity concerns directly.

Multi-turn protocols. Where the underlying paradigm is inherently sequential (hindsight bias, feeling-of-knowing, Iowa Gambling) we preserve the multi-turn structure rather than collapsing to a single-shot probe.

4. From Profiles to Theory

A degradation profile is a function $\rho_b : \mathcal{D}_b \to [0, 1]$ mapping the difficulty axis of benchmark b to expected accuracy. A candidate theory of the underlying capability predicts a *shape* for ρ_b: log-linear decay for CDI under a noisy channel model of attention, a learning curve of the form $1 - e^{-k/\tau}$ for ALGIn under a Bayesian-induction account, a step at the order-3 boundary for RBT under a working-memory–bounded ToM account, and so on. The benchmark then becomes a falsifier: deviations from the predicted shape are diagnostic, in a way that a single accuracy number cannot be.

This is the role we see the battery playing in the theory–benchmark virtuous cycle that motivates the workshop: each benchmark is a parametric probe whose output is rich enough for theoretical models to commit to a prediction, and structured enough that mismatches between prediction and observation localise the defect to a specific construct.

Worked example: an RBT prediction. A working-memory bounded account of recursive ToM (Ullman, 2023) predicts a sharp drop between order 2 and order 3 (a "ToM cliff") with a shallower slope on either side; a uniform-cost search account predicts a smooth log-linear decay across orders 1–4. RBT items at orders 1–4 with anti-heuristic controls let us fit both shapes to a model's ρ_{RBT} and reject one. The same template lets us repeat the test across model scales, asking whether the cliff softens with scale (consistent with capacity arguments) or remains fixed (consistent with an architectural bound). Analogous shape-vs-shape contrasts apply to the other four benchmarks: log-linear vs. step decay in CDI, exponential vs. power-law learning in ALGIn, and so on.

Cross-benchmark profile. Each model is summarised by a five-curve *cognitive profile* $(\rho_{\mathrm{CDI}}, \rho_{\mathrm{ALGIn}}, \rho_{\mathrm{ECUU}}, \rho_{\mathrm{DRO}}, \rho_{\mathrm{RBT}})$, with each curve scalarised at need (e.g. slope and intercept of the degradation fit) for compact cross-model tables. Profiles support two kinds of analysis: (i) qualitative comparison across models and scales, and (ii) correlational analysis across the 1,138 sub-task types, which lets us ask whether the five construct families are empirically separable in LLMs or collapse onto fewer latent factors.

Table 1. Battery summary. Item and sub-task counts refer to the full procedurally-generated battery; the released v0.1 CSV snapshots contain a smaller sample for inspection.

Benchmark	Items	Sub-tasks	Score	Difficulty axis
CDI	4,938	229	bool	0–6 distractors
ALGIn	5,585	223	bool	2/4/8/12 examples
ECUU	4,975	226	float	10 unknow. tiers
DRO	4,976	230	bool	6 prepotent levels
RBT	4,916	230	bool	orders 1–4
Total	25,390	1,138		

5. Human Baselines and Calibration

For each benchmark we draw on three sources of human baseline data, in decreasing order of directness. First, where an item type is a direct adaptation of a published paradigm (e.g. Sally–Anne for RBT order 1, Stroop for DRO inhibition, Hart's feeling-of-knowing for ECUU), the original human distribution transfers and is used as-is. Second, for procedurally generated items that share a parameter fam-

ily with a published study but vary along an untested axis (e.g. RBT at recursion orders 3–4), we report the published baseline at the closest tested point and flag the extrapolation. Third, for entirely new families with no plausible mapping (ALGIn at higher rule counts, ECUU on the unknowable tier, prepotent suppression in DRO), we plan a small ($n \approx 30$ per cell) human pilot, documented in the v0.2 release, rather than imputing a baseline.

We do not claim "a human baseline exists for every item." We do claim that every item is traceable to a paradigm with a documented baseline at *some* parameter setting, and that the gap, where it exists, is made explicit. The pilot will use a single procedurally generated item per cell per participant (so the generator, not a fixed instance, is what is benchmarked), recruit through an online platform with a comprehension-check pre-screen, and report per-cell mean accuracy with bootstrap 95% CIs alongside the model curves; the protocol will be pre-registered with the v0.2 release.

6. Limitations and What This Paper Is Not

This paper introduces the battery and its design rationale; it does *not* report model evaluation, calibration, or correlational results. Those are deferred to a companion release because the fairness and informativeness of cross-model comparisons depend sensitively on prompt protocol, scoring procedure, and contamination auditing, and we want the design decisions to be public and critiqued before any leaderboard is locked in. Three open questions follow directly.

Construct validity. We have argued that, e.g., CDI probes selective attention rather than long-context handling, but the argument is currently structural (controlled lures, position manipulation, length-matched controls). Empirical separation (e.g. showing that CDI scores at fixed context length still vary with distractor count, and that the residual after partialling out long-context performance correlates with other attention probes) is required and is part of the v0.2 plan.

Calibration against humans. The plan in Section 5 commits us to pilot data for the non-derivable cells. Until that is collected, the human-comparability claim is one of *paradigm continuity*, not of point-matched baselines.

Sufficiency of five constructs. A factor analysis over model scores on the 1,138 sub-task types will tell us whether the five-way partition is empirically supported or whether some constructs (e.g. executive function and metacognition) collapse for current LLMs. We treat this as an empirical question for the next release, not as an axiom of the paper.

7. Conclusion

We have described a five-benchmark cognitive battery for foundation models, comprising 25,390 procedurally generated items across 1,138 sub-task types. Each benchmark is built so its output is a degradation profile along a controlled difficulty axis, which is both the kind of object a theoretical account can commit to a prediction about and the kind of object whose deviations are diagnostic. The benchmarks are released as instruments; the calibration, model evaluation, and construct-validity analyses they support are the subject of a companion release.

References

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

Broadbent, D. E. *Perception and Communication*. Pergamon Press, 1958.

Cattell, R. B. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1):1–22, 1963.

Cherry, E. C. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, 1953.

Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Fischhoff, B. Hindsight $\neq$ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3):288–299, 1975.

Flavell, J. H. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10):906–911, 1979.

Hart, J. T. Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56(4):208–216, 1965.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. 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, 2000.

Rozenblit, L. and Keil, F. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5):521–562, 2002.

Saffran, J. R., Aslin, R. N., and Newport, E. L. Statistical learning by 8-month-old infants. *Science*, 274(5294): 1926–1928, 1996.

Simons, D. J. and Chabris, C. F. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, 1999.

Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.