A TEXT-FREE PRESCHOOL MATHEMATICS BENCHMARK: WHAT ITS SCORES DO AND DO NOT MEASURE

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a procedurally generated benchmark for preschool mathematics — counting, comparing, adding and subtracting with counters, pattern continuation, conservation of number and equal sharing — designed for a cascaded agent of small models rather than a single large one. The benchmark is rendered entirely without text: identity, size, position, container contents, holding state, episode phase and counting boards are encoded as colour, geometry and layout, so that no task can be solved by reading. Where a family admits a non-verbal answer the agent must demonstrate the result by acting, and both cascade stages are *educated* on corpora generated from the same distribution, shipped with the benchmark. We evaluate 21 configurations over 608 evaluation episodes and find that the training method, not the architecture, accounts for the largest gap in the results: full-parameter training of a 2B front end reaches 0.809 where low-rank training of the same model reaches 0.480, and a quarter of the education data with full-parameter training beats all of it with low-rank training. The front end bounds overall performance, with the back end contributing a family-specific competence profile that does not scale monotonically. We then audit the benchmark itself, computing for each family the score attainable by an agent that neither perceives nor reasons. Sixteen of 19 families carry usable signal, two are saturated, and one has none: conservation of number reports a peak success rate of 0.53 against a floor of 0.500, so its apparent difficulty is an answer prior being collected, and no configuration clears its own constant strategy by more than one item. We argue that this floor computation belongs in the benchmark's release gate rather than in post-hoc analysis, and that for families with skewed answers the headline metric must be balanced accuracy reported against chance.

1 INTRODUCTION

A benchmark score is a claim about what an agent can do. Two design decisions decide whether that claim is true, and both are easy to get wrong in ways that leave the score looking healthy.

The first is what the task actually requires. A benchmark that renders identifiers, counts or labels into the image is not measuring perception: an agent that has learned to read solves it. The failure is silent because the score is high. The second is whether the tasks can discriminate at all. A family whose ground-truth answers are skewed is partly solved by the answer prior: an agent that always emits the majority answer and never opens the image collects that share of the score, and a family whose trajectories saturate is solved by everyone and ranks no one. Again the failure is silent.

This paper reports a benchmark built under the first constraint and audited under the second. It targets preschool mathematics — counting, comparing, adding and subtracting with counters, pattern continuation, conservation of number, equal sharing — for a cascade of small models rather than a single large one. Every task is procedurally generated, and the scene is rendered to a single image containing *no text whatsoever*: identity, size, position, container contents, holding state, episode phase and, where relevant, a counting board are all encoded as colour, geometry and layout. Where a family admits a non-verbal answer, the agent demonstrates the result by acting.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The agent is the two-stage cascade of Figure 2, with both stages *educated* on data generated from the same task distribution. The decomposition is the experimental instrument, not an implementation convenience: swapping one stage while holding the other fixed separates perception failure from planning failure, and makes its difficulty claims falsifiable.

We evaluate 21 configurations, spanning three vision architectures, six back ends, two training methods, three corpus fractions and a two-device setup for the 8B stage. The results decompose where difficulty lives rather than ranking the models against one another.

The audit is the paper's second contribution and produced a negative result about our own benchmark: of 19 families, 16 carry usable signal, 2 are saturated and 1 has none. Conservation of number peaks at 0.531 against a floor of 0.500, and no configuration clears its own constant strategy there by more than one item in 32; the aggregate had looked like a hard family being cracked.

- A procedurally generated, fully text-free benchmark for preschool mathematics: 19 families over four levels, 608 evaluation and 38,912 training tasks, leakage controlled by a three-stage guard, and both education corpora shipped alongside it.
- A 21-configuration study that separates method from architecture from data volume, showing that low-rank training — not architecture or data — accounts for the largest single gap in the results, and that the front end bounds overall performance.
- A floor-auditing procedure for benchmark families, applied as a negative result on one of our own: the floor computation belongs in the release gate rather than in post-hoc analysis, and for skewed families the headline metric must be balanced accuracy against the chance line rather than a raw rate that the prior inflates.
- Family-level evidence that the back end carries a family-specific competence profile: one family is solvable only by the two largest back ends, while another ranks the smallest first, in both cases independently of the front end.

2 Related work

Embodied and vision-language benchmarks. Embodied benchmarks have moved from hand-designed scene-and-task suites to procedurally generated, physically simulated environments (Shridhar et al., 2020; Deitke et al., 2022), and toward scale and photorealism (Savva et al., 2019). We take from this line procedural generation with explicit leakage control and shipping the generator with the benchmark, and we draw our families from a documented developmental sequence (Sarnecka & Carey, 2008) rather than a manipulation inventory.

Educational agents and preschool mathematics. Whether a model can be *taught* a competence it does not acquire zero-shot connects early-childhood education — what it requires of an agent that questions and scaffolds rather than presents answers (Xu et al., 2024) — with the arithmetic limits of language models, which fail in ways diagnostic of representation rather than reasoning capacity (Nogueira et al., 2021). Our benchmark sits between them: its arithmetic is trivial symbolically, and the difficulty lies in obtaining quantities visually and manipulating them without a symbol system. The design follows the developmental account in which non-symbolic quantity representation precedes symbolic arithmetic (Halberda et al., 2008; Wynn, 1992), conservation of number is a milestone rather than a computation (Houdé et al., 2011), and number-word meanings are acquired in a fixed sequence (Sarnecka & Lee, 2009); two of our three added families are computational analogues of classical paradigms there. Children's over-trust of conversational agents (Van Brummelen et al., 2023) informs Section 4.

Modular and cascaded small-model architectures. Decomposing a system into perception and planning modules is an old idea (Andreas et al., 2016; Ahn et al., 2022; Jiang et al., 2021); our cascade is a straightforward instance. The question is not whether decomposition helps in general but what it makes measurable: the front-end/back-end attribution used throughout Section 5. Our results bear on low-rank adaptation against full-parameter training (Hu et al., 2021; Biderman et al., 2024) and on the reach of sub-2B models (Lu et al., 2024): our best configuration uses a 2B front end and a 1.7B back end. A contrasting line removes the interface altogether: end-to-end vision-language-action models discretise continuous commands into a fixed token vocabulary and co-train

one network to emit them alongside language (Brohan et al., 2023; Kim et al., 2024), trading an opaque action map for generalisation from web-scale pretraining.

Benchmark validity: base rates and discriminative power. Reporting benchmark scores without an account of what a non-expert strategy would score has a documented history (Ma et al., 2024; McCoy et al., 2019); shortcut solutions are a recognised failure mode, and for agent benchmarks how a run is scored is itself contested (Agarwal et al., 2021; Peng et al., 2025). Contamination between the data-generation process and the evaluation set is a documented failure mode (Draoui et al., 2026; Fund et al., 2025), and whether a capability appears to exist can itself be an artefact of the metric chosen to measure it (Schaeffer et al., 2023): the choice of metric is part of the capability claim, not neutral reporting.

Our contribution relative to this line is procedural: computing each family's floor and enforcing it at release time, which found a family of ours that had been read as a hard capability (Section 6). The nearest 2024–2026 benchmarks (Liu et al., 2026; Chen et al., 2026; Li et al., 2025; Weng et al., 2025) test models, not children. CoreCognition's own limitations name two gaps we address: a "complementary evaluation format . . . to include non-linguistic or purely visual models" left to future work, and a controlled probe of shortcuts whose "meticulous nature limits scalability". Our text-free constraint (Section 3.3) takes the first, and the release-time floor audit (Section 6) the second.

3 THE BENCHMARK

3.1 DESIGN CONSTRAINTS

Three constraints fix the design, and each rules out a shortcut that would otherwise make the benchmark easier than it looks.

Non-verbal. The tasks are preschool mathematics: counting, comparing, adding and subtracting with counters, continuing a pattern, conserving number, sharing equally. Crucially the *answer modality* is non-verbal wherever the family allows it: in manipulation families the agent demonstrates the result by moving counters rather than naming it, and where a verbal answer is unavoidable the instruction is written for a listener who cannot yet read. A benchmark accepting a spoken numeral as the answer measures something easier than what a five-year-old can do.

Pure visual. The rendered frame contains no text at all — no identifiers, region labels, numerals, legends or badges — so every task-relevant fact must be recovered from geometry and colour; the renderer issues zero text-drawing calls and a test asserts it. Text in the image converts a visual task into a reading task.

Educated, not prompted. Both stages are trained on data generated from the same task distribution. The shortcut this rules out is not prompting as such but prompting *in place of* perception: a pretrained model asked to describe a scene it cannot ground will answer from its language prior, and a high score would then measure the prior rather than the image. What the ablation tests is what *education* buys inside such a cascade: we compare trained configurations against one another, not against a large general model, so our results do not bound what scale alone would achieve.

3.2 TASK FAMILIES

The released version contains **19 families** over four levels, **608** evaluation tasks (19×32) and **38,912** training tasks (19×2048). Families are grouped by the capability they target: L1 perception, L2 manipulation, L3 reasoning (including the three added families below), and L4 memory and recovery. The evaluation set is stratified L1 128 / L2 96 / L3 224 / L4 160; the full list of families with their capabilities and answer modalities is in Appendix A.

Three families were added to move the set closer to preschool mathematics itself. `F_SEQUENCE_QUANTITY` presents piles following an arithmetic or doubling progression and asks for the next; its instruction contains no numeral and the largest candidate is never the answer, so

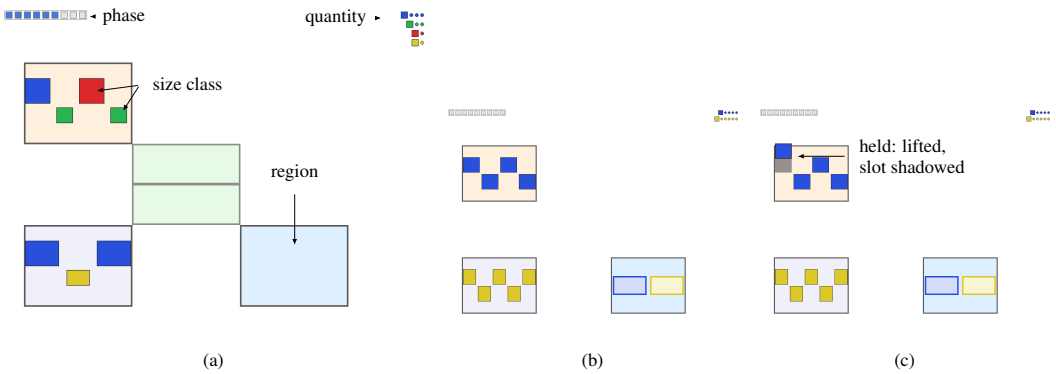

Figure 1: **Visual encoding.** All three panels come from the frozen benchmark renderer, which draws no text at all; the arrows and labels are annotations we have added. (a) A cross-episode rule task. (b,c) One sorting task before and after `reach+grasp` on `blue_block_01`: the held block lifts out of its slot and leaves a translucent shadow.

"pick the biggest" cannot succeed. `F_CONSERVE_NUMBER` is a computational analogue of the classical conservation task: one pile is compared with two smaller piles of the same total, and the answer is not recoverable from any single visible count. `F_EQUAL_SHARE` requires an exact division of T counters into $B \in \{2, 3\}$ baskets. Their strata fix the answer prior by construction (16 equal, 8 left-greater, 8 right-greater), which removes the sampling accident but not the prior — a distinction Section 6 shows to separate a hard family from an uninformative one.

3.3 Visual encoding

The scene is rendered to a single 800×600 image; Figure 1 shows the vocabulary and Appendix B lists every cue with its pixel parameters. Two choices are load-bearing: families whose semantics a visible total would give away use blocks rather than counters, so no dot board is drawn, making the encoding part of the task design rather than a rendering detail; and the phase pip bar makes a memory task's distractor phase distinguishable from its answer phase without text, which is what lets the non-verbal memory families exist at all in the released set.

The renderer must match byte-for-byte between education-data generation and evaluation, so we treat it as a frozen, versioned artefact, served to the agent as one.

3.4 Generation and leakage control

Tasks are generated procedurally from per-family templates with parameterised phrasing. Because training and evaluation come from the same generator, the benchmark is exposed to the usual contamination failure, and we guard it in three places. First, the phrase banks are hard-partitioned per family: a module-load assertion fails if any training phrase for a family appears verbatim in that family's evaluation bank (151 training, 113 evaluation phrases, disjoint across all 19 families). Second, generation applies a containment guard: a candidate evaluation instruction is rejected if a training instruction of at least eight characters occurs verbatim inside it, or it inside that instruction. The unit compared is the whole formatted instruction (4,325 unique strings across the pool), so a rejection requires one complete sentence embedded in another, not a shared fragment. Third, a release gate re-checks the artefacts: evaluation text against all training text, evaluation instances and layouts against the training pool, and a check that no instance signature is duplicated within the training pool itself. It flags none of the 608 evaluation instructions under the containment rule.

The gate reports 38,912 training and 608 evaluation tasks against 33,912 tasks of prior-version training data, with disjoint phrase banks and no duplicated instances. It also reports 192 evaluation *layouts* that also occur in training, which we keep: a layout is the geometric arrangement of regions, and what must not repeat is the instance — the specific objects, colours and quantities. Sharing layouts is deliberate, so a gap reflects the task rather than a shift in image statistics.

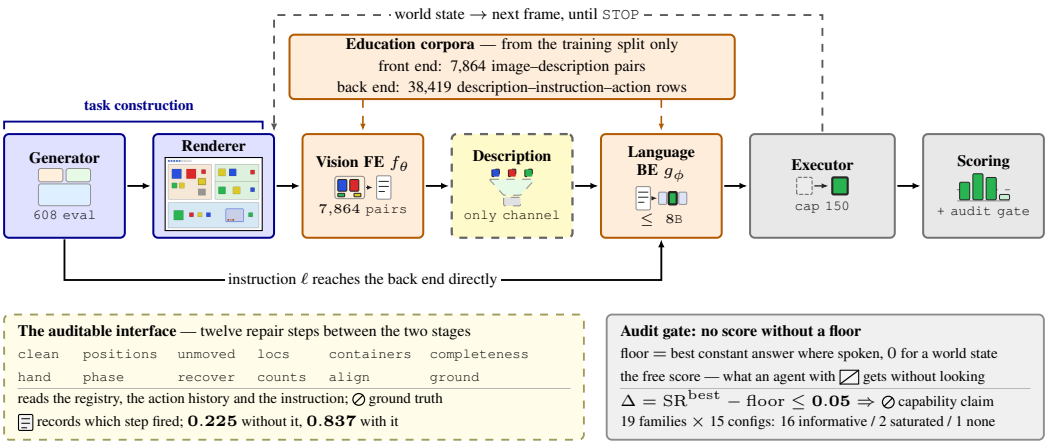

Figure 2: **The cascade.** Blue is task construction; the dashed *education corpora*, drawn from the training split only, train the two sub-8B stages; yellow is the auditable interface. The picture in the Renderer box is a schematic — 800 × 600, with no text at all. The instruction reaches the back end directly, so the front end never sees it during the episode, and scoring ends at the audit gate of Section 6.

3.5 Education data

The benchmark ships with the data used to educate the agent. For the vision front end, a balanced sample of the training pool — 24 tasks per family, hence 456 tasks covering all 19 — is replayed through the executor and rendered at every step, giving 7,864 image–description pairs: one per recorded world state, not one per pool task. The same replay supplies the language back end with 38,419 rows over a wider sample, including several phrasings of one task. A back-end row is a description–instruction–action triple: the scene field the back end is given at evaluation is embedded in every training row, written by the same function. Both corpora are drawn from the training split only, and the generator ships with the benchmark so they can be regenerated exactly.

The training pool is far larger than any single run uses, so the ablation draws stratified subsets of 25% and 50% of the corpus, separating the effect of *how* a stage is trained from that of *how much* data it sees.

4 Protocol

4.1 Task formalism

Each task is a tuple $(W_0, \ell, \mathcal{G})$: an initial world state, an instruction and ground truth. The world is a tabletop rendered to a single image (Section 3.3) containing objects, their regions and, for counting tasks, a dot board. Ground truth has three parts: `final_state`, the predicates that must hold when the episode ends; `answer`, the verbal answer, present only for verbal families; and `action_program`, a canonical solving sequence used as the reference policy for path efficiency (SPL).

4.2 Action interface

The agent acts through eight primitives — `reach`, `grasp`, `move`, `put_down`, `look_at`, `point_at`, `wait`, `say` — and terminates with `STOP`, within a cap of 150 steps. Each is a symbolic command over object identifiers, applied by a deterministic executor; identifiers are not visible in the image, so the back end infers them from the description. An infeasible action (grasping out of reach, placing into an occupied slot) leaves the state unchanged, and the fourth rejected action ends the episode. The planner is shown its last three actions but not whether they succeeded, so a failed step is recoverable only by re-reading the scene — which is what the recovery families require.

4.3 CASCADED AGENT

A *vision front end* f_θ consumes the rendered frame and emits a structured textual description; a *language back end* g_ϕ consumes that description with the instruction and emits the next action. Both stages are small ($\leq$ 8B) and both are *educated* on corpora generated from the benchmark (Section 3.5). The description protocol is deliberately narrow — a fixed field layout over regions, objects and attributes — so the interface between the stages is auditable.

What the interface passes through. The transcript is not handed over verbatim. It passes through a repair layer of twelve named steps that consults the executor's registry and the agent's own action history first: entities the world does not contain are dropped and an invented identifier is rewritten onto the real object whose colour and region match what the model itself reported; the position and held-state fields of objects the agent has moved are overwritten; target containers the front end omitted are re-attached; the counter ledger is overwritten with the counts the executor recorded when the agent looked at a board; and the location list is reordered by the colour the instruction names. The layer is part of the protocol, applied identically to every configuration in Table 1, and shipped behind a single environment variable. It never reads the ground truth for the current frame and cannot add information the agent has not obtained by acting — but the scores we report measure the cascade *with* this repair, not the two learned stages alone. Section 5 quantifies what it contributes.

4.4 METRICS

An episode yields a trajectory $\tau = (a_1, \ldots, a_T)$ and terminal world W_T; $L = |\tau|$ and o is the optimal length. We report five quantities.

Success rate. $\mathrm{SR} = 1$ iff no predicate in `final_state` is violated *and* either the task has no required answer or the final `say` matches `answer` — both required.

Path-weighted success. $\mathrm{SPL} = \mathrm{SR} \cdot o/\max(L, o)$, with $o = \max(1, |\texttt{action_program}|)$ in raw primitives so that o and L share a unit; phased memory tasks additionally include the mandated distractor waits and one `look_at` per recalled selector, and unphased verbal tasks take $o = 2$.

Episodic question answering. Where the answer is spoken rather than achieved, $\mathrm{EQA} = 1$ iff the `say` matches `answer`; undefined for the rest, where the check is vacuous.

Coverage. The fraction of ground-truth target objects touched by `reach` or `grasp` — separating an agent that fails after manipulating the right objects from one that never found them.

Five-year-old criterion. With $r = (1 - |L - o|/o) \cdot \mathrm{SR}$ if $L \leq 2o$ and $r = 0$ otherwise, $5\mathrm{YC} = \mathbb{1}[r \geq 0.5]$ — deliberately blunt, a coarse indicator only.

4.5 VERIFICATION

Verification is symbolic and complete: predicates over the final world are checked exactly against `final_state`, and answers are compared after canonicalisation. No learned judge is used, so no metric depends on one model's opinion of another. Every number in Section 5 is a mean over 608 episodes.

5 EXPERIMENTS

5.1 SETUP

Except where stated, all configurations are evaluated on the 608 evaluation episodes of Section 3, with identical protocol, prompt and renderer, and every stage is trained for exactly one epoch (optimiser and rank settings in Appendix F). Each number is a mean over 608 episodes, 32 per family, unless noted; we read only differences far above that resolution. One constraint shapes the study: a stage's full-parameter training must fit a single 32 GB device. Our V100s have no native bfloat16 support, so the bfloat16-pretrained models (InternVL3, Gemma-3) are trained in float32 — float16 yields a zero loss and NaN gradients on this hardware. Gemma-3-4B full-parameter was thereby

Table 1: Six-group matrix, 608 evaluation episodes each; *OK* is the number of those episodes solved. Both stages are trained for one epoch on the education data of Section 3.5.

	Front end	Front method	Back end	OK	SR
G1	Qwen3-VL-2B	full	Qwen3-1.7B	508	**0.8355**
G2	Qwen3-VL-4B	full	Qwen3-4B	505	0.8306
G3	Qwen3-VL-8B	LoRA	Qwen3-8B	386	0.6349
G4	Gemma-3-4B	LoRA	Phi-4-mini	360	0.5921
G5	InternVL3-2B	LoRA	Gemma-3-1B	240	0.3947
G6	InternVL3-1B	LoRA	DeepSeek-R1-1.5B	137	0.2253

ruled out at a 34.4 GB footprint, and we sharded the 8B front end alone to test whether the regime saturates at the top of the range.

5.2 The matrix

Table 1 pairs six front ends of three vision architectures with a comparable back end. It is not an architecture ranking: architecture is confounded with training method, since the leading Qwen3-VL front ends were trained full-parameter and the rest with LoRA — and the matched comparisons below show that reading the InternVL3 rows as an architectural weakness is wrong.

5.3 Findings

We refer throughout to the 15 ablation configurations of Appendix D.

Method, not architecture, is the binding constraint. Matched comparisons isolate each factor. Changing the front-end architecture while holding method and back end fixed is worth at most 0.076 (InternVL3-1B LoRA $0.4523 \rightarrow$ Qwen3-VL-2B LoRA 0.5280). Changing only the method of one fixed architecture is worth 0.329: A8 is InternVL3-2B trained full-parameter in float32, scoring 0.8092 against 0.4803 for A3, the same architecture under LoRA, landing within 0.026 of G1 — so the matrix's low InternVL3 scores are a low-rank artefact, not an architectural ceiling. The converse does not hold: with the full-parameter Qwen3-VL-2B front end fixed, three back ends span 0.074 (0.8273, 0.8355, 0.7615); at the 8B front end an 8B and a 1.7B back end differ by 0.005; and A9, the weakest front end given the strongest back end, reaches only 0.5444.

Method dominates data. At a fixed architecture, back end and corpus only the training method changes: G1 and `B1b` share the Qwen3-VL-2B front end, the Qwen3-1.7B back end and the identical corpus, differing only in how the front end is trained — full-parameter 0.8355 against LoRA 0.5280. Within full-parameter training the data curve over that same corpus is $0.6250 \rightarrow 0.7105 \rightarrow 0.8355$ at 25%, 50% and 100%, so doubling the data buys at most $+0.125$: a rank-16 update given all of the data lands below full-parameter training on a quarter of it. Both arms are one epoch, so this compares at equal optimisation budget, not at convergence.

Saturation at 8B. C8 trains the 8B front end full-parameter across two devices and reaches 0.8520; swapping in the 1.7B back end gives 0.8470, three episodes fewer. Against G3 — the same 8B model under LoRA — full-parameter training is worth $+0.217$, but scaling the front end from 2B to 8B buys only $+0.017$ over G1. The regime is bounded by the front end at roughly 0.85.

5.4 Where the failures are

A family solved only by the back end. `F_SUB_COUNTERS` is solvable if and only if the back end is Qwen3-8B or Phi-4-mini. Across all 21 configurations there is no counterexample: back ends at 1.5B, 1.7B, 4B and Gemma-3-1B score exactly zero, while Phi-4-mini scores 0.84–0.91 and Qwen3-8B 0.94–1.00. The cleanest demonstration is a controlled flip: the same C8 front end yields 0.9375 with the 8B back end and 0.0000 with the 1.7B one, and the reverse pairing — Qwen3-1.7B planning with the strongest front end — also scores zero. The failure is not perceptual; the same

description supports success when a different planner reads it. The 1.7B back end scores 0.91 on F_ADD_COUNTERS, so the deficit is specific to subtraction, not arithmetic as such.

Two families no configuration solves. The two added families that require transforming a perceived quantity rather than reporting it are the only two nothing solves: F_SEQUENCE_QUANTITY peaks at 0.3125, F_CONSERVE_NUMBER at 0.531 — the latter, as Section 6 shows, not evidence of capability at all.

F_SEQUENCE_QUANTITY is the odd one: its score is a function of the back end alone, taking six values across the 21 configurations and no others — 0.3125 for Qwen3-1.7B (nine of them), 0.25 for Phi-4-mini, 0.125 for Qwen3-4B, 0.094 for Gemma-3-1B, 0.0625 for Qwen3-8B, 0 for DeepSeek-R1-1.5B. The collapse is exact: configurations sharing a back end emit the same action sequence item by item, so thirteen configurations spanning six front ends, both training methods and three corpus fractions are indistinguishable from one another. The ordering is also inverted — smallest first — though only the 1.7B-against-8B gap survives a paired test ($p = 0.039$).

Together with F_SUB_COUNTERS, which has the opposite ordering, that is a family-specific competence profile rather than a scale factor. The invariance also says the obstacle sits in the planning stage, which is where it belongs, and it is not a capacity limit either, since configurations with more capacity do worse rather than better. Whether the obstacle is how a transformed value is represented, the mapping from a count onto an object-level action vocabulary, or the family's size at 32 items is not something our design separates; we report the invariance and leave the mechanism open.

Weak pairs are recoverable. With one weak front end held fixed (InternVL3-1B with LoRA), stronger back ends give $0.2253 \rightarrow 0.4523 \rightarrow 0.5444$ for DeepSeek-R1-1.5B, Qwen3-1.7B and Phi-4-mini: the back end's leverage is largest where the front end is weakest, the opposite of the pattern at 8B. The secondary metrics (Appendix E) show SPL tracking SR to within 0.010 throughout.

What the interface contributes. Disabling the repair layer of Section 4 — one environment variable, everything else held fixed on a re-run of G1's checkpoints — costs the cascade most of its score: 137 of the 608 episodes solved against 509, winning 372 paired episodes and losing none ($p < 10^{-110}$), with twelve of the 19 families scoring exactly zero. What survives carries its answer in the raw transcript; what needs rewriting vanishes. A frontier model pair with no education and one in-context protocol example solves 31 of 128 episodes of an earlier stratified subset (SR 0.2422, one run; disjoint from Table 1), tying an untrained 2B cascade on those episodes — the same failure shape as the unrepaired cascade here (0.2253): free-form prose meeting a back end that expects a fixed layout. The interface fails before the reasoning does.

6 What the scores actually measure

A benchmark's scores are only interpretable if its tasks can discriminate between agents. A family on which every configuration scores the same, or on which a strategy that neither perceives nor reasons matches the best cascade, contributes nothing to a ranking — but contributes to the aggregate, invisibly. We therefore audit the released benchmark and ship the audit with it.

6.1 Floor and headroom

For each family we compute a *floor*: the score achievable with neither perception nor reasoning. Where the answer is communicated rather than achieved — by speech, or by pointing at an object — the floor is the best constant strategy, $\text{floor} = \max_a \Pr[\text{gt} = a]$, which an agent that never looks at the image attains; where it is a world state the floor is zero, exact for a multi-step family and understated for a single-action one, where a fixed choice of target has a nonzero chance of being right. The error runs in the direction of keeping a family rather than discarding it. The quantity of interest is the headroom $\Delta = \text{SR}^{\text{best}} - \text{floor}$: the score above what is available for free. A family with $\Delta \leq 0.05$ cannot separate the configurations we evaluate and must not support a claim about capability, however high its raw score.

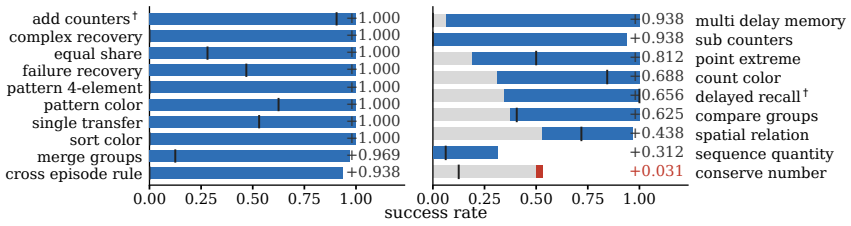

Figure 3: **Discrimination audit, all 19 families, ranked by headroom (left column, then right).** Grey is the *floor* — the score available to an agent that never looks at the image — and the coloured segment is the headroom Δ above it; the black tick is the lowest score any of the 15 configurations attains. Blue families separate the configurations, the red one does not, and † marks the two that are saturated. Family names are printed here in expanded lower case; the tables use the released identifiers (`F_PATTERN_4ELEMENT` for *pattern 4-element*).

The result, over the 15 ablation configurations, is that **16 of the 19 families have usable headroom, 2 are saturated, and 1 carries no signal at all** (Figure 3; Appendix C gives the per-family table). We discuss the three failure modes, none visible from the aggregate score.

Saturated families. `F_DELAYED_RECALL` is solved perfectly by every configuration (spread 0.00): outgrown, not defective. `F_ADD_COUNTERS` spans only 0.09 (from 0.91 to 1.00), so almost all of its variance is noise. Neither can rank configurations any more, and the remedy is to raise their difficulty, not to report them as evidence of competence.

A family with no signal. `F_CONSERVE_NUMBER` is the interesting case, and the reason we consider the audit a contribution rather than a formality. Its raw success rate reaches 0.531, but its floor is 0.500: the ground truth is "equal" on 16 of 32 items, "right greater" on 8 and "left greater" on 8, so an agent that always answers "equal" without looking already scores 0.50. The headroom is $+0.031$, below the 0.05 threshold for no signal.

Appendix Table 5 decomposes the family: of the 15 ablation configurations only two clear the constant strategy at all, each by a single item in 32, while two land exactly on it and eleven fall below. The 0.531 headline comes from those two — the same 8B front end with two back ends — which answer "equal" on 21 of 32 items: the answer prior, not the agent.

Consequence: report balanced accuracy, and gate on headroom. Two changes follow. For a skewed family the headline metric must be balanced accuracy against the chance line, since the raw success rate is dominated by the prior; and the floor computation belongs in the release gate, so that a family below threshold fails the build the way a duplicated instance does. Re-stratifying `F_CONSERVE_NUMBER` would fix the prior but not the family: under a uniform distribution the floor drops to 0.333 and the best balanced accuracy is still 0.479. Balancing is necessary and not sufficient — conservation of number, for this cascade, is beyond reach. The audit states which families can support conclusions, not whether the benchmark is valid as a whole, and no conclusion in Section 5 rests on `F_CONSERVE_NUMBER`.

7 Conclusion

We built a text-free benchmark for preschool mathematics, educated a cascade of small models on it, and audited it before trusting its scores. The audit changed what we could conclude: one family's apparent difficulty was $+0.031$ over a strategy that never looks at the image, and two were solved by every configuration. Difficulty lives in the method, not the architecture: a rank-16 update cannot install the description protocol even given the entire corpus, while full-parameter training succeeds with a quarter of it. Given a competent method the front end bounds performance, and the back end carries a family-specific competence profile, not a scale factor. Three caveats bound this: an equal-budget rather than converged comparison, one configuration the hardware ruled out (Section 5), and one model family — our families are computational analogues of developmental paradigms, not a study of children.

AI USE STATEMENT

We used generative AI tools for tasks that require disclosure under the ICLR 2027 policy. Specifically, we used them to **implement methods** — the cascaded agent, the scene renderer, the training and evaluation harness, and the audit script; to **design and provide feedback on research methodology and experiments** — the configuration matrix, the ablation tiers, and the discriminative-power audit; to **generate synthetic data sets** — the procedurally generated task families and the education corpora; and to **interpret results** — including the front-end/back-end attributions in Section 5 and the family-level verdicts in Section 6. We have not used generative AI tools for the collection of human-subject data, and that task is not applicable to this work, as the benchmark contains no human participants.

Additionally, we used generative AI tools for tasks with recommended disclosure: creating and editing software code, suggesting experimental parameters, creating scientific figures, drafting and editing parts of the text for readability, structuring the paper, and formatting references.

We have reviewed all AI-assisted work. Every number reported in this paper was produced by executing the released evaluation harness on the released benchmark, not by a language model; the audit in Section 6 is likewise computed by a script over evaluation outputs. One of the AI-assisted components produced a result we subsequently retracted: an early automated check reported three artefact-size mismatches in the released tree, which manual inspection showed to be false positives caused by comparing hand-edited files against their unedited sources. The affected component was corrected and the check re-run before any number in this paper was produced. The human authors verified the benchmark's leakage gate output, inspected the per-family audit figures, and confirmed that no claim in this paper rests on `F_CONSERVE_NUMBER`. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The benchmark involves no human participants and no personal data: all scenes are synthetic and all agents are models. Two considerations nonetheless apply. First, the work is motivated by early-childhood mathematics education, and benchmarks in that space invite a specific misuse — reading a score as evidence that a system is fit to teach children. Nothing in this paper supports that inference: the agents we study manipulate counters and report quantities, they do not interact with learners, and our best configuration fails two families outright. The educational framing is a source of task design, not a deployment claim. Second, the benchmark's difficulty is deliberately calibrated to a young child's competence, which means a high score indicates a narrow, non-verbal competence rather than general capability; the audit in Section 6 exists partly to make that boundary explicit, by reporting which families can and cannot support conclusions about an agent at all.

REPRODUCIBILITY STATEMENT

We release the benchmark, the generation code, both education corpora, the evaluation harness and the audit script. Every configuration in Sections 5 and 6 is identified by its front end, back end, training method and data fraction; the remaining hyperparameters are given in Section 5. Task generation is seeded and deterministic, and the released evaluation set is fixed, so a re-run reproduces the reported numbers exactly. The renderer is a frozen artefact of the benchmark, and the byte-identity requirement between the training-time and evaluation-time renderer (Section 3.3) is asserted by a test. The one result we cannot release in full is a configuration ruled out on hardware grounds (A7), for which no score exists; we report it as a limitation rather than a result. Per-family values for all configurations are included with the artefact.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the study, to the construction and release of the benchmark, to the design and execution of the experiments, and to the analysis and writing of the manuscript. The benchmark's task families and generation pipeline, the scene renderer and evaluation protocol, the training and evaluation harness, and the discriminative-power audit were each developed collaboratively, with individual authors leading different components; the corresponding artefacts are released with the paper.

ACKNOWLEDGMENTS

We thank the reviewers for their time. The experiments in this paper were run on a small number of institutional GPUs; the single-device memory constraint that shaped the configuration matrix and the one configuration ruled out on hardware grounds (Section 5) both reflect that setup. Named funding acknowledgements are deferred to the camera-ready version.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. arXiv preprint, 2021. URL https://doi.org/10.48550/arXiv.2108.13264.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint, 2022. URL https://doi.org/10.48550/arXiv.2204.01691.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39–48. IEEE, 2016. doi: 10.1109/CVPR.2016.12. URL https://doi.org/10.1109/CVPR.2016.12.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. arXiv preprint, 2024. URL https://doi.org/10.48550/arXiv.2405.09673.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint, 2023. URL https://doi.org/10.48550/arXiv.2307.15818.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Y. charles, Yiping Bao, Yuantao Fan, Guopeng Li, Haiyang Shen, Xuanzhong Chen, Wendong Xu, Shuzheng Si, Zefan Cai, Wenhao Chai, Ziqi Huang, Fangfu Liu, Tianyu Liu, Baobao Chang, Ming Wu, Xiaobo Hu, Kaiyuan Chen, Yixin Ren, Yang Liu, Yuan Gong, and Kuan Li. BabyVision: Visual reasoning beyond language. arXiv preprint, 2026. URL https://doi.org/10.48550/arXiv.2601.06521.

Matt Deitke, Eli Vanderbilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems 35*, pp. 5982–5994. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-0433. URL https://doi.org/10.52202/068431-0433.

Taha Draoui, Mohamed Wiem Mkaouer, and Christian Newman. Improving data leakage detection in machine learning notebooks through static slicing and structured LLM prompts. *Proceedings of the ACM on Software Engineering*, 3(FSE):4371–4392, 2026. doi: 10.1145/3808199. URL https://doi.org/10.1145/3808199.

Fraida Fund, Mohamed Saeed, Shaivi Malik, and Kyrillos Ishak. Learning from irreproducibility: Introducing data leakage case studies for machine learning education. In *Proceedings of the 3rd ACM Conference on Reproducibility and Replicability*, pp. 224–228. ACM, 2025. doi: 10.1145/3736731.3746153. URL https://doi.org/10.1145/3736731.3746153.

Justin Halberda, Michèle M. M. Mazzocco, and Lisa Feigenson. Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213):665–668, 2008. doi: 10.1038/nature07246. URL https://doi.org/10.1038/nature07246.

Olivier Houdé, Arlette Pineau, Gaëlle Leroux, Nicolas Poirel, Guy Perchey, Céline Lanoë, Amélie Lubin, Marie-Renée Turbelin, Sandrine Rossi, Grégory Simon, Nicolas Delcroix, Franck Lamberton, Mathieu Vigneau, Gabriel Wisniewski, Jean-René Vicet, and Bernard Mazoyer. Functional magnetic resonance imaging study of Piaget's conservation-of-number task in preschool and school-age children: A neo-piagetian approach. *Journal of Experimental Child Psychology*, 110(3):332–346, 2011. doi: 10.1016/j.jecp.2011.04.008. URL https://doi.org/10.1016/j.jecp.2011.04.008.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. arXiv preprint, 2021. URL https://doi.org/10.48550/arXiv.2106.09685.

Yifeng Jiang, Michelle Guo, Jiangshan Li, Ioannis Exarchos, Jiajun Wu, and C. Karen Liu. DASH: Modularized human manipulation simulation with vision and language for embodied AI. In *The ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, pp. 1–12. ACM, 2021. doi: 10.1145/3475946.3480950. URL https://doi.org/10.1145/3475946.3480950.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. arXiv preprint, 2024. URL https://doi.org/10.48550/arXiv.2406.09246.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D. Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, and Hokin Deng. Core knowledge deficits in multi-modal language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34379–34409. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25p.html.

Jingping Liu, Xueyan Wu, Hanxuan Chen, Ziyan Liu, Zhangquan Chen, Ronghao Chen, and Huacan Wang. Easy for children, hard for AI: The limits of multimodal LLMs in early childhood learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32078–32086, 2026. doi: 10.1609/aaai.v40i38.40479. URL https://doi.org/10.1609/aaai.v40i38.40479.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D. Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights. arXiv preprint, 2024. URL https://doi.org/10.48550/arXiv.2409.15790.

Chong Ma, Lin Zhao, Yuzhong Chen, Lei Guo, Tuo Zhang, Xintao Hu, Dinggang Shen, Xi Jiang, and Tianming Liu. Rectify ViT shortcut learning by visual saliency. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):18013–18025, 2024. doi: 10.1109/TNNLS.2023.3310531. URL https://doi.org/10.1109/TNNLS.2023.3310531.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1334. URL https://doi.org/10.18653/v1/P19-1334.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Investigating the limitations of transformers with simple arithmetic tasks. arXiv preprint, 2021. URL https://arxiv.org/abs/2102.13019.

Jierui Peng, Yanyan Zhang, Yicheng Duan, Tuo Liang, Vipin Chaudhary, and Yu Yin. NEBULA: Do we evaluate vision-language-action agents correctly? arXiv preprint, 2025. URL https://doi.org/10.48550/arXiv.2510.16263.

Barbara W. Sarnecka and Susan Carey. How counting represents number: What children must learn and when they learn it. *Cognition*, 108(3):662–674, 2008. doi: 10.1016/j.cognition.2008.05.007. URL https://doi.org/10.1016/j.cognition.2008.05.007.

Barbara W. Sarnecka and Michael D. Lee. Levels of number knowledge during early childhood. *Journal of Experimental Child Psychology*, 103(3):325–337, 2009. doi: 10.1016/j.jecp.2009.02.007. URL https://doi.org/10.1016/j.jecp.2009.02.007.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019. doi: 10.1109/ICCV.2019.00943. URL https://doi.org/10.1109/ICCV.2019.00943.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? arXiv preprint, 2023. URL https://doi.org/10.48550/arXiv.2304.15004.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10737–10746. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01075. URL https://doi.org/10.1109/CVPR42600.2020.01075.

Jessica Van Brummelen, Maura Kelleher, Mingyan Claire Tian, and Nghi Nguyen. What do children and parents want and perceive in conversational agents? towards transparent, trustworthy, democratized agents. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, pp. 187–197. ACM, 2023. doi: 10.1145/3585088.3589353. URL https://doi.org/10.1145/3585088.3589353.

Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. VisNumBench: Evaluating number sense of multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3830–3840, 2025. doi: 10.1109/iccv51701.2025.00365. arXiv:2503.14939.

Karen Wynn. Addition and subtraction by human infants. *Nature*, 358(6389):749–750, 1992. doi: 10.1038/358749a0. URL https://doi.org/10.1038/358749a0.

Ying Xu, Kunlei He, Julian Levine, Daniel Ritchie, Zexuan Pan, Andres Bustamante, and Mark Warschauer. Artificial intelligence enhances children's science learning from television shows. *Journal of Educational Psychology*, 116(7):1071–1092, 2024. doi: 10.1037/edu0000889. URL https://doi.org/10.1037/edu0000889.

A TASK FAMILIES

Table 2: All 19 task families, named by their identifiers in the released artefact. "Ans." is the answer modality: *act* = the answer is a world state the agent must achieve, *say* = the answer is communicated, by speech or by pointing.

Family	L	Capability	Ans.
F_COUNT_COLOR	L1	count objects of one colour	say
F_COMPARE_GROUPS	L1	compare two piles	say
F_POINT_EXTREME	L1	identify the largest object	say
F_SPATIAL_RELATION	L1	report left/right relation	say
F_SINGLE_TRANSFER	L2	grasp and place one object	act
F_SORT_COLOR	L2	sort objects by colour	act
F_MERGE_GROUPS	L2	merge two piles	act
F_ADD_COUNTERS	L3	represent a sum with counters	act
F_SUB_COUNTERS	L3	represent a difference	act
F_PATTERN_COLOR	L3	continue an alternating pattern	act
F_PATTERN_4ELEMENT	L3	continue a four-slot pattern	act
F_SEQUENCE_QUANTITY	L3	extend a quantity progression	act
F_CONSERVE_NUMBER	L3	judge conservation of number	say
F_EQUAL_SHARE	L3	distribute equally into baskets	act
F_DELAYED_RECALL	L4	recall a count after distraction	say
F_MULTI_DELAY_MEMORY	L4	recall two colours and counts	say
F_FAILURE_RECOVERY	L4	recover from a failed action	act
F_COMPLEX_RECOVERY	L4	recover from multiple failures	act
F_CROSS_EPISODE_RULE	L4	reuse a rule across episodes	act

B VISUAL ENCODING REFERENCE

Table 3: The complete visual vocabulary. The renderer issues no text-drawing calls; every cue below is carried by colour, geometry or layout.

Semantic	Encoding
Object identity, type	Coloured solid square; four-colour palette, with a hash-seeded fallback for unnamed colours
Size class	Box scaling, $1.35\times$ (*big*) against $0.72\times$ (*small*); the anti-overlap clamp then bounds the drawn boxes at 46–72 against 30–42 pixels
Position in region	Placement among the region's anchor slots (west, east, middle, north, south, centre)
Region extent	Grey-bordered rectangle with a light slot tint
Container	Filled shape with coloured outline; contents as a row of chips
Held object	Lifted out of its slot, translucent grey shadow left behind
Episode phase	Nine-cell pip bar, top-left, fixed phase order
Counting board	Per-colour dot rows, top-right; drawn only when the world has counters

C FULL DISCRIMINATIVE-POWER AUDIT

Table 4: All 19 families, ordered by headroom, over the 15 ablation configurations. Floor = score available with neither perception nor reasoning (best constant strategy for *say* families, zero for *act* families); $\Delta = \mathrm{SR}^{\mathrm{best}} - \mathrm{floor}$.

Family	Type	Floor	Min	Max	Δ	Verdict
F_ADD_COUNTERS	act	0.000	0.906	1.000	+1.000	saturated
F_COMPLEX_RECOVERY	act	0.000	0.000	1.000	+1.000	informative
F_EQUAL_SHARE	act	0.000	0.281	1.000	+1.000	informative
F_FAILURE_RECOVERY	act	0.000	0.469	1.000	+1.000	informative
F_PATTERN_4ELEMENT	act	0.000	0.000	1.000	+1.000	informative
F_PATTERN_COLOR	act	0.000	0.625	1.000	+1.000	informative
F_SINGLE_TRANSFER	act	0.000	0.531	1.000	+1.000	informative
F_SORT_COLOR	act	0.000	0.000	1.000	+1.000	informative
F_MERGE_GROUPS	act	0.000	0.125	0.969	+0.969	informative
F_CROSS_EPISODE_RULE	act	0.000	0.000	0.938	+0.938	informative
F_MULTI_DELAY_MEMORY	say	0.062	0.000	1.000	+0.938	informative
F_SUB_COUNTERS	act	0.000	0.000	0.938	+0.938	informative
F_POINT_EXTREME	say	0.188	0.500	1.000	+0.812	informative
F_COUNT_COLOR	say	0.312	0.844	1.000	+0.688	informative
F_DELAYED_RECALL	say	0.344	1.000	1.000	+0.656	saturated
F_COMPARE_GROUPS	say	0.375	0.406	1.000	+0.625	informative
F_SPATIAL_RELATION	say	0.531	0.719	0.969	+0.438	informative
F_SEQUENCE_QUANTITY	act	0.000	0.062	0.312	+0.312	informative
F_CONSERVE_NUMBER	say	0.500	0.125	0.531	+0.031	**no signal**

Table 5: F_CONSERVE_NUMBER, ten of the 15 ablation configurations. Columns 3–5 are per-class recall for the three answers; the last is balanced accuracy (mean per-class recall). Random guessing scores 0.333 and the best constant strategy scores 0.500 raw (16 of 32 items). The five omitted configurations (B1a, B2a–d) also fall below it, at 4, 7, 8, 10 and 11 items. Configurations are named as in Table 6.

Configuration	Raw SR	Equal	Right	Left	Balanced
C8a, C8b	0.531	0.69	0.62	0.12	**0.479**
A8	0.500	0.69	0.50	0.12	0.438
A6	0.500	0.75	0.38	0.12	0.417
A2	0.281	0.06	0.00	1.00	0.354
A9	0.250	0.00	0.00	1.00	0.333
A3 / A4	0.219	0.00	0.00	0.88	0.292
A5	0.219	0.38	0.00	0.12	0.167
B1b	0.125	0.12	0.00	0.25	0.125

D FULL ABLATION RESULTS

Table 6: All 15 ablation configurations at 608 evaluation episodes each; *OK* is the number of those episodes solved. The A-tier holds the back end fixed at Qwen3-1.7B except where noted; the B-tier retrains a front end on 25% or 50% of the education corpus; the C-tier trains the 8B front end full-parameter across two devices. Each row is a distinct trained artefact, and each identifier occurs exactly once in the paper.

ID	Front end	Back end	OK	SR
A2	Gemma-3-4B LoRA	Qwen3-1.7B	319	0.5247
A3	InternVL3-2B LoRA	Qwen3-1.7B	292	0.4803
A4	InternVL3-1B LoRA	Qwen3-1.7B	275	0.4523
A5	Qwen3-VL-2B full	Phi-4-mini	503	0.8273
A6	Qwen3-VL-2B full	Gemma-3-1B	463	0.7615
A7	Gemma-3-4B full	—	—	ruled out (34.4 GB)
A8	InternVL3-2B full	Qwen3-1.7B	492	0.8092
A9	InternVL3-1B LoRA	Phi-4-mini	331	0.5444
B1a	Qwen3-VL-4B LoRA (100%)	Qwen3-4B	327	0.5378
B1b	Qwen3-VL-2B LoRA (100%)	Qwen3-1.7B	321	0.5280
B2a	Qwen3-VL-2B full (25%)	Qwen3-1.7B	380	0.6250
B2b	Qwen3-VL-2B full (50%)	Qwen3-1.7B	432	0.7105
B2c	Qwen3-VL-4B full (25%)	Qwen3-4B	396	0.6513
B2d	Qwen3-VL-4B full (50%)	Qwen3-4B	493	0.8109
C8a	Qwen3-VL-8B full	Qwen3-8B	518	0.8520
C8b	Qwen3-VL-8B full	Qwen3-1.7B	515	0.8470

E SECONDARY METRICS

Table 7: The secondary metrics of Section 4.4 for all 15 ablation configurations, aggregated from the same episodes as Table 6. EQA is defined for the seven *say* families of Table 4 — those whose answer is communicated rather than achieved — and is averaged over those 224 of 608 episodes; the other four metrics are means over all 608.

Configuration	SR	SPL	Coverage	EQA	5YC
C8a	0.852	0.850	0.572	0.848	0.852
C8b	0.847	0.845	0.574	0.929	0.847
A5	0.827	0.827	0.578	0.879	0.827
B2d	0.811	0.809	0.575	0.879	0.811
A8	0.809	0.808	0.558	0.897	0.809
A6	0.762	0.752	0.552	0.893	0.753
B2b	0.711	0.707	0.543	0.795	0.709
B2c	0.651	0.651	0.522	0.692	0.651
B2a	0.625	0.625	0.501	0.696	0.625
A9	0.544	0.544	0.359	0.754	0.544
B1a	0.538	0.536	0.432	0.692	0.538
B1b	0.528	0.527	0.428	0.692	0.528
A2	0.525	0.525	0.410	0.688	0.525
A3	0.480	0.471	0.349	0.638	0.467
A4	0.452	0.451	0.352	0.585	0.452

Two readings of the table matter. SPL tracks SR to within 0.010 everywhere, so configurations that solve a task solve it close to the reference path: path efficiency is not a bottleneck that the success rate conceals. Coverage separates configurations that SR alone does not. A9 scores 0.016 above B1b in aggregate and 0.069 below it in coverage; restricted to the twelve action families, where the metric is meaningful, the coverage gap widens to 0.110 while their success rates there are within 0.010. The two reach comparable aggregates by different routes: A9 gains on the verbal families

(0.753 against 0.692), which is the back end's contribution, while `B1b` finds the ground-truth objects more often on the manipulation families, which is the front end's.

F TRAINING DETAILS

Every stage is trained for exactly one epoch with a constant learning rate — no warmup and no decay — and no family receives a schedule of its own. Batch size and sequence cap are set by stage, and the settings below are read off the manifests each run wrote, not restated from memory. Vision stages use an effective batch size of 16, as a 2×8 or 1×16 microbatch–accumulation split, except the sharded 8B front end, whose two-device configuration runs at 8; the language back end uses 32 when full-parameter and 16 when low-rank. Sequence length is capped at 1024 tokens for vision stages — 2048 for the InternVL3 front ends — and at 2048 for the back end. The optimiser is AdamW-8bit for every full-parameter stage, and AdamW for every low-rank stage except the Qwen3-VL front ends in low-rank mode, which also used AdamW-8bit. Learning rates are 2×10^{-5} for every vision stage, and 2×10^{-5} (10^{-4}) for a full-parameter (low-rank) back end. LoRA uses rank 16, $\alpha = 32$ and dropout 0.05. The seed is 42 for every vision stage and 20260803 for the back end. Inference is served by vLLM on two V100-32GB devices.

One epoch is a protocol choice rather than a budget limit, and it does not explain the method gap. On one fixed low-rank arm — the 2B front end of `B1b`, the same corpus — extending training to three and five epochs raises success from 0.5280 to 0.6151 and 0.6645 (paired McNemar $p = 5 \times 10^{-5}$ and $p = 0.0095$), yet five epochs remain 0.171 below `G1`, the one-epoch full-parameter arm on the same corpus, on 120 episodes won against 16 lost ($p = 7 \times 10^{-21}$). More optimisation budget on the same method does not close the gap that changing the method does.

G COMPUTE

All training and inference ran on NVIDIA V100-32 GB devices, one device per stage, with a single exception: the 8B front end is the only configuration that does not fit one device, and it is sharded across two. That sharding is fully sharded data parallelism, and gradient synchronisation is issued at every micro-step rather than accumulated across micro-batches — the accumulating path grew peak memory by about 3 GB per step on this stack and did not finish a run, while per-micro-step synchronisation costs extra communication and trains the same model. For that front end the embedding and language-model head are frozen, so its trainable set matches the one the low-rank arm it is compared against trains. The dtype is set by the hardware rather than chosen (Section 5).

Device memory, not time, is what decided which configurations entered the matrix, so this report gives the topology and no wall-clock figures.

H RELEASE AND VERSIONING

The artefact is versioned by release date. Every number in this paper is computed on the splits of snapshot `v2026-09-11`: the `v1.4` evaluation set of 608 episodes and the `v1.4` training pool of 38,912, released together with the generation code, both education corpora, the evaluation harness and the audit script. Licence terms and a public location are deferred to the camera-ready version.

I PER-FAMILY RESULTS

Per-family success rates for every configuration, and the audit's per-configuration decomposition of `F_CONSERVE_NUMBER`, are available in the released artefact (Appendix H). The audit of Appendix C ships with it as a script that recomputes every floor and headroom from the released ground truth and the stored per-episode outcomes, independently of the numbers printed here.