Pre-registered gates for small models trained on one laptop: what passed, what failed, and what the readings show

maxgfr · 20 November 2026 · PDF · How to cite

Technical report, not peer-reviewed

This report has not been peer-reviewed, nor reviewed by anyone outside the project. Every number in it is copied from a page or a file of the repository, and each one links to its evidence. Where a number is computed here (a count, or the difference of two published numbers), the text says so.

Abstract

skycmd is an open-source project that trained small models for typed decisions, tool routing, secret detection, end-of-turn detection, forecasting and memory on one Apple M5 laptop, and runs them in a browser from the files their numbers were measured on. It asks what such models can do when every claim is judged against a bar, a gate, written down in advance. From s5 on, each gated experiment had a pre-registration committed before the computation it governs (24 documents), its test was read once, every change was a dated amendment, and a model that missed its gate was published with its file. Of the 20 published model files, 15 were judged by a gate fixed before training: 5 passed and 10 failed; the 5 earlier files had no gate. One pass is a second look at test records already read, and another is a system designed after a test was read and judged on fresh records. Nine more versions stopped without a model, eight of them pre-registered. Four observations cut across lines: early-exit thresholds fitted on held-out questions failed when checked by playing (two models); readings on data built differently from training limited three passes and decided two failures, while most failures came on the line's own test; one gate could be passed by a constant answer, and another caught a tie only because a lexical baseline was added before training; and controls that hide or rename the input exposed a training shortcut. Pre-registrations, result files and model files are public.

---

1. Introduction

A *decision model* receives a state (a game frame, a short text, the last seconds of audio) and one or more typed questions, and returns a probability for every allowed option in one forward pass, without generating text. The question types are `choice` (one of several options), `score` (an ordered level) and `noul` (a probabilistic yes or no) (`docs/research.md`). This format, called "System One" by the models that introduced it in September 2026, is served by models of 17M parameters (oscar-1-17m [1]) to 27B; the open models in it that the project found start from a pretrained backbone of at least 17M parameters (What we bring, section 1).

skycmd asks how far below that size such a model can go, and what it can be trusted with, when it is trained from scratch on one laptop. Its first answer was s3-pico, 445,345 ternary parameters in a 251,104-

byte file, which plays four control games at the level of an exact oracle (manifest, What we bring, section 1). The later lines, s4 to s6, carried the same engine to questions in free text, to sound, to tool lists given at run time, to secrets in code, to time series and to a fast-weight memory.

This report contributes three things.

1. **Artefacts.** Model files from 142,912 bytes (s6-cast-pico) to 3,692,288 bytes (the two files of s6-route7-nano) for the s6 family, and up to 20,684,640 bytes for s5-lite, each published with its card, its verdict and a demo that runs it in the browser from the file the numbers were measured on (manifest).
2. **A protocol applied to every line from s5 on**, with the failures published as results: a pre-registration frozen before the computation it governs, a test read once, dated amendments, a checkpoint rule fixed in advance, and a second look named as such (section 3.3).
3. **Observations across lines** (section 5), each stated with its evidence and its limits. None of them was pre-registered as a test; they are patterns in the published readings.

Its claim to being a scientific result rests on the protocol, not on a leaderboard number. Several of the models are weak; the point is that how weak they are, and where, was fixed in advance and is checkable.

---

2. Related work

This section cites only work that the repository already credits (Sources); each reference was checked against its arXiv or project page. The literature search behind the project was two research passes, not a systematic review (What we bring), and the project credits no prior work on pre-registration in machine learning, so this report does not survey it.

Decision models. TypeSafe's Jev (a closed API, 15 September 2026) launched the format; Laya [2] (421M parameters), Julia-1 [3] (144M) and oscar-1-17m [1] are open models in it, and llama.cpp serves it at `/v1/systemone` (`docs/research.md`). The project's s5 line builds on two papers about the format: one shows that typed decision models follow the option's name rather than its definition [4], and another makes the answer independent of the options' order with a block-causal attention [5].

Architecture. A small shared block looped several times with an answer after each loop follows TRM [6] and Ouro [7]. One token per measured value follows FT-Transformer [8] and FoNE [9]. Ternary weights trained in follow BitNet b1.58 [10], and their behaviour at small scale is studied by Nielsen et al. [11].

Calibration, early exit and imitation. Temperature scaling and the expected calibration error (ECE) with 15 equal-width bins are those of Guo et al. [12]. Early-exit thresholds with a guarantee come from Conformal Risk Control [13], in the early-exit setting of CALM [14] and Fast yet Safe [15]; Learn then Test [16] underlies the s5 escalation experiment. Imitation that holds up in closed loop uses DAgger [17].

Optimisers. The pre-registered ablation (section 4.9) was designed knowing that gains over AdamW shrink once AdamW is tuned as carefully [18] and that, at equal time, AdamW can beat Muon on small models [19].

Benchmarks, teachers and baselines. The end-of-turn line is judged against Smart Turn v3.2 [20] on its test set and on LiveKit's eot-bench [21]; the forecasters on GIFT-Eval [22] and Chronos-ZS [23], against TinyCast [24]; the secrets verifier on Samsung's CredData [25]; the routers against BM25, the static

embedding potion-base-2M [26] and all-MiniLM-L6-v2 [27], with fresh tools from Toucan [28] and the When2Call set [29]. The memory line tests the "sleep" passes of Lee et al. [30].

---

3. Setting and protocol

3.1 The architecture in one page

The decision models share one engine, written in Rust, with the same graph for training (through mlx-rs on Apple Silicon) and for inference (natively and in WebAssembly) (What we bring, sections 2 and 6).

- **Two streams with a cache.** The question and its options form a static stream, encoded once and cached; each frame encodes only the state. In a game, the same question is asked at every frame, so only the state is new.
- **Options that never see each other.** No option attends to another, so an answer cannot depend on the options' order; this holds by construction, and on the s2 models permuting the options moved no probability by more than 3e-6 on the CPU (`BENCH.md` , section "s2 training runs"). In the s5 iso format, each option sees the state, the question and itself only, at shared positions, and the option keys are never rendered, so renaming them cannot change an answer either. These properties are checked by exact tests to 1e-6, natively and in WebAssembly; they are tests, not experimental results (Results, s5).
- **One token per measured value.** A pair such as `vy -3` becomes a single token whose embedding carries the number. In s3-pico this nearly halves the state: 9 tokens instead of 16 for a Flappy frame (`BENCH.md` , section "Inference").
- **Depth from recursion.** One shared state block is applied several times, with an exit (an answer) after each pass. s3-pico has 1 block applied 4 times, width 96 and a 1,024-token vocabulary; s6-nano has 2 blocks applied twice each, width 144, a 2,048-token vocabulary and windows of 512 tokens for the state and 160 for the question (`BENCH.md` , section "skycmd-s3"; `docs/s6-nano2-study.md` §1). An early exit answers at the first pass whose top probability exceeds a threshold τ; τ = 1 means full depth.
- **Ternary weights trained in.** Weights take the values –1, 0 and +1 times a scale (absmean quantization), trained with a straight-through estimator after a float warm-up (quantization-aware training, QAT). The files store 2-bit weights.

The end-of-turn models of s6-ear v2 and v3 use a dedicated graph (`skycmd-turn` : 4 blocks, attention pooling, 40 log-mel bands plus pitch, voicing and energy per 10 ms hop), and the forecasters a patch encoder on the same blocks (`skycmd-cast`) (manifest; Results, s6-ear2-nano; Results, s6-cast-pico).

3.2 The training setting

Every model was trained on one Apple M5 laptop with 16 GB of memory, the machine its owner works on (`docs/big-run-report.md`). A *brake* lets training run hard when the machine is idle and gently when it is in use, on battery or hot: a daemon writes a duty cycle and a work-and-rest pattern to one file, and the training loop rests between steps. The rests stay out of the run's clock, so every budget is counted in **working minutes** and computes the same thing in every mode; only the wall-clock duration changes (`docs/idle-training.md`). The project ran one budgeted GPU run at a time.

The training data are public sets under their own licences, game oracles written in Rust, and labels from open-weight teachers. The teacher of s6-nano and s6-micro, GLM-5.3-flash (open weights, MIT licence), labelled 747,077 of the 747,198 selected training questions through its provider's API; the 121 others were refused by the provider's moderation (`docs/big-run-report.md` §2 and §7). The end-of-turn student was distilled from Smart Turn v3.2 (BSD-2-Clause) and the forecasters from TinyCast (Apache-2.0) (Results, `docs/big-run-report.md` §5). Every source and its licence is listed in `DATA_LICENSES.md`.

3.3 The protocol

The rules below are those of `AGENTS.md`, section "Pre-registered experiments", applied from s5 on; `docs/README.md` indexes every pre-registration with its status.

- **Pre-registration.** Before a model is trained, a file in `docs/` fixes the data and its splits, the model and its budget, the test it will be read on, the gate (a set of conditions that must all hold), what is reported without being gated, and what a failure would mean. It is committed before any computation it governs. There are 24 such files (a count made here of `docs/*prereg*.md`).
- **A test read once.** A test set is opened once, by the step the pre-registration names. For s5, the evaluation tool refused the test split of a frozen set unless the pre-registration listed the set's hash, and logged every pass over it (`docs/s5-prereg.md`). In later lines the evaluation entry points refuse the test until a marker file (`TEST-OPEN` , or `FRESH-OPEN` for fresh files) is created after the validation steps; for s6-route-nano and s6-route that file was created under the owner's standing delegation, not by the owner's hand, and the result blocks say so (`docs/s6-nano2-prereg.md` , `docs/s6-nano3-prereg.md`).
- **Dated amendments.** Any change is an amendment, dated and committed before the computation it concerns; it says what had been seen when it was written and what does not change. Amendments include technical ones (a memory cap, a resumed run) and decisions of the owner (publishing failed files, closing a line).
- **The checkpoint rule is fixed in advance.** For example the last DAgger round at full depth for s6-nano, or the exponential moving average of the weights for s6-cast-pico and s6-ear3-nano; no intermediate checkpoint is chosen by looking at it (Results).
- **No rerun to see.** Only a technical failure justifies a new run, and it is recorded. Two examples: s6-nano's decision training was stopped by the memory cap at step 16,020 and resumed from its checkpoint at step 15,573; s6-micro v2's died on a file-access error and was resumed from step 61,051 (`BENCH.md` , section "skycmd-s6").
- **Failures are published as results.** Since the owner's decision of 15 November 2026, a model that fails its gate is published with its file, a card that opens with the failure and a manifest entry marked `gate: failed` ; since 17 November 2026 each has a demo page (a Hugging Face Space) (`AGENTS.md`).

Four terms are used throughout this report.

- **Second look.** A system designed after a test was read. When it is judged on the same records, its result is not an independent test: this is the case of s6-micro v2 (a pass) and s6-secrets v3 (a failure). When it is judged on records built afterwards, the test is new but the design is not blind to the earlier reading: this is the case of s6-route, whose fusion was first seen on s6-route-nano's already-read test. All three are called second looks wherever they appear below.

- **In-sample for design.** A test set whose readings shaped a later version's design choices, though never its weights: Smart Turn's test clips for s6-ear2-nano and s6-ear3-nano, GIFT-Eval for s6-cast2-pico (Results, limits).
- **Reported, not gated.** A number the pre-registration announced but that decides nothing.
- **Exploratory.** A reading made outside any pre-registration: the closed-loop comparison of s2-nano's thresholds, and the studies and diagnoses written after a gate from files already read. None of them changes a verdict, and each is named as such where it is used.

---

4. Results

4.1 Counts

The counts below are made here from the manifest of published files (`manifest.json`), the status table of Results and the index of `docs/README.md` .

- **Published model files: 20.** Gate passed: 5 (s6-nano, s6-micro v2, s6-secrets-pico, s6-ear2-nano, s6-route). Gate failed: 10 (s4-micro, s6-micro-v1, s6-ear-nano, s6-ear3-nano, s6-cast-pico, s6-cast2-pico, s6-secrets-verify-pico, s6-route-nano, s6-route7-nano, s6-sleep3-nano). No gate: 5 (s1-pico, s2-nano, s2-nano-doom and s3-pico, built before the project fixed gates, and s5-lite, a longer run declared outside the s5 bets for the demo).
- **Of the 15 gated files**, 14 had a pre-registration document frozen before training; s4-micro had a gate fixed before training and recorded in `BENCH.md` , with no pre-registration document (`docs/README.md`).
- **Second looks: 3.** s6-micro v2 (passed, same records), s6-route (passed, fresh records, design after a test reading) and s6-secrets v3 (failed, same records; not a new file).
- **The status table of Results** has 17 rows: 5 passed, 11 failed, 1 stopped by its pilot. Its 11 failures are 9 published s6 files, the s5 bets and the System 1 / System 2 router of LFM3-1.7B, which trained no weights. The table covers the pre-registered lines from s5 on; s4-micro, whose gate has no pre-registration document, is named in a sentence under it.
- **Versions closed without a model: 9.** Eight had a pre-registration. Seven of them stopped at a check or rule it named: s6-sleep v1 (its pilot), v2 (a CPU check) and v4 (two checks on training sequences); s6-route v4, v5 and v6 (data checks, before any training); the LFM3 fine-tuning (no method fitted the machine's memory bound). The eighth, four flexible-weights experiments under one pre-registration, was closed without running after a literature review. A ninth, a possible s6-secrets v4, was closed by a diagnosis and had no pre-registration.

Table 1 lists the gated files and what they do.

Table 1. The 15 published files judged by a gate fixed before training. Parameters and bytes from the manifest and Results.

model	what it does	parameters	bytes	verdict
s6-nano	four games and typed questions in free text	1,486,225	715,136	passed
s6-micro (v2)	the same, larger	5,587,201	2,420,032	passed, second look
s6-micro-v1	the same, larger	5,587,201	2,420,032	failed
s4-micro	typed questions in free text	5,587,201	2,419,456	failed
s6-secrets-pico	is a secret exposed in this text?	445,345	250,272	passed (synthetic bench)
s6-secrets-verify-pico	is this rule candidate a real credential?	445,345	249,984	failed
s6-ear-nano	end of turn and keywords from audio	1,532,449	899,200	failed
s6-ear2-nano	has the speaker finished?	1,306,178	631,936	passed
s6-ear3-nano	the same, retrained on conversations	1,306,178	631,936	failed
s6-cast-pico	probabilistic forecast, 64 steps	369,792	142,912	failed
s6-cast2-pico	the same, with a periodicity front end	377,568	148,512	failed
s6-route-nano	pick a tool from a list given at run time	5,587,201	2,418,912	failed
s6-route	the same file fused with a static embedding	5,587,201 and an embedding	3,692,224	passed, second look
s6-route7-nano	the router retrained on rebuilt data	5,587,201 and an embedding	3,692,288	failed
s6-sleep3-nano	s6-nano with a fast-weight memory	1,611,505	752,416	failed

4.2 Every gate, as read

Table 2 gives each gate's main bar, the measurement and the verdict, with the caveat that belongs next to the number. Intervals are 95% bootstrap intervals over test items, as published.

Table 2. The gates of the 15 gated files.

model	main bar	measured	verdict	stated with the number
s6-nano	`typed_decisions_bench` > 0.398 (majority class); ECE ≤ 0.08; no game more than 5% below s3-pico	0.480; ECE 0.0479 [0.0416, 0.0564]; the oracle's score in all four games	passed	a constant answer per task passes the bench bar: each question's most common training answer scores 0.4815 (study)
s6-micro v2	bench ≥ 0.45; ECE ≤ 0.08; games	0.553; ECE 0.0588 [0.0526, 0.0687]	passed	second look: designed after v1's ECE reading, judged on the same test records
s6-micro-v1	as v2	0.547; ECE 0.0821 [0.0746, 0.0900]	failed	the gate is read on the point value; the interval holds 0.08 (report §6)
s4-micro	bench > 0.55	0.343	failed	gate fixed before training, no pre-registration document
s6-secrets-pico	secret > 0.977 and service > 0.787 (a rule scanner on the same texts)	0.9990; 0.9795	passed	the bench is the project's own synthetic generator; on 719 real files with no secret it flags 59.8% at p ≥ 0.9 (real text)

model	main bar	measured	verdict	stated with the number
s6-secrets-verify-pico	line F1 ≥ best rule scanner + 0.15 = 0.8832 (G1), and six more gates	0.8561; over the best rule scanner +0.1229 [0.0413, 0.3613]	failed (G1; six of seven hold)	the `UUID` rows were removed by a choice made before any test file was read; v3, a second look, reads 0.8477
s6-ear-nano	end of turn ≥ Smart Turn v3.2 (0.9263); keywords ≥ 0.90; zero-shot keywords ≥ 0.20 in English and French	0.5681; 0.8885; 0.2100; 0.1330	failed (three of four)	the bar was first read as 0.8518 through a padding bug, then corrected; the verdict does not change
s6-ear2-nano	accuracy ≥ 0.85 and ≥ 0.9260 − 0.08; false cut-off ≤ 0.20; END recall ≥ 0.80; ≤ 5 ms in WebAssembly	0.8993 [0.8962, 0.9027]; 0.0847; 0.8627; 2.39 ms	passed	Smart Turn's test is partly in-sample for its design; on eot-bench its AUC is 0.7299, its teacher's 0.7838
s6-ear3-nano	E1: AUC(v3) − AUC(v2) on eot-bench, 0.2 s into each pause, ≥ +0.02 with a lower bound > 0	−0.0252 [−0.0336, −0.0167] (0.7197 against 0.7448)	failed (E1, E2, E3; A1, A2, G3 hold)	designed after v2's eot-bench reading; label cross-check amended after a measurement on five training sessions
s6-cast-pico	GIFT-Eval nGMASE ≤ 0.838 and nWQL ≤ 0.587	0.9471; 0.6646	failed	
s6-cast2-pico	the same (G1); MASE under the seasonal naive's at every term (G2)	0.8847; 0.6190; G2 holds	failed (G1)	GIFT-Eval is partly in-sample for its design
s6-route-nano	above BM25 on unseen tools (G1); ≤ 2 ms in WebAssembly (G4)	0.8395 against 0.8500 (−0.0105 [−0.0265, +0.0055]); 38.9 ms	failed (G1, G4)	
s6-route	above BM25 with a lower bound > 0; ≥ MiniLM − 0.03 (0.8748); ECE; ≤ 60 ms	0.8868 against 0.8169 (+0.0698 [0.0455, 0.0952]); ECE 0.0162; 30.9 ms	passed	second look (fresh records); the latency bar was set after v2 was measured; without the three largest clusters +0.0202 [−0.0062, +0.0481]
s6-route7-nano	above BM25 with a lower bound > 0 (G1); ECE ≤ 0.05 (G3); at chance with the request hidden (G6)	−0.0065 [−0.0235, +0.0120]; 0.1845; 0.1969 against at most 0.1929	failed (G1, G2, G3, G6; G4, G5 hold)	designed after v4 to v6's data tables
s6-sleep3-nano	H0: memory ≥ control + 0.10 on `recall`; H1: sleep gain lower bound > 0; H2: games at most 2% under the control	+0.2452 [+0.2370, +0.2532]; +0.0030 [−0.0010, +0.0068]; Flappy 15.87 against 71.05	failed (H1, H2 and the s6-nano gate; H0, H3 hold)	three training seeds, pooled

Table 3 gives the experiments whose gate or check was read without producing a published gated file.

Table 3. Gated readings and stops without a published gated file.

experiment	bar	measured	outcome
s5, three bets	bet 1: ECE of the iso format at most 0.01 above the usual format's	0.088 against 0.057 (+0.031 [+0.006, +0.054]); accuracy 0.674 against 0.429	failed (all three bets)
LFM3-1.7B System 1 / System 2	H2: accuracy within 0.02 of always thinking; H3: at most half its tokens	0.786 against 0.796; 1,032 tokens against at most 0.5 × 958	failed (H3); no weights trained
s6-secrets v3	the seven gates of v2 on rules and verifier together	G1 0.8477 against 0.8832	failed, second look
s6-sleep v1	pilot: memory ≥ control + 0.10 on `recall`	0.2460 against 0.2650	stopped, no nano run

experiment	bar	measured	outcome
s6-sleep v2	CPU check C4: the memory is used	300 training steps did not learn to use it	stopped before any GPU run
s6-sleep v4	K3 and K1: `recall` ≥ 0.95 on training sequences, on both seeds	K3 0.2579 and 0.2546; K1 0.8442 and 1.0000	closed, no model
s6-route v4	data checks: no rule that ignores the request beats chance by its tolerance	the shortest name right on 0.4219 against a chance of 0.3721	stopped before training
s6-route v5	the same checks on surface-matched distractors	+0.028 to +0.076 above chance in eight cells; *none* readable on 0.4410 against 0.20	stopped before training
s6-route v6	its stop rule on When2Call's slates	*none* readable from the options' spread on 0.5792 against a share of 0.20	stopped before training
LFM3 fine-tuning	a LoRA method that ends an optimizer step under the memory bound	none did	closed, nothing trained
flexible weights	four experiments	not run	closed after a literature review

The rest of this section gives, line by line, the numbers the tables cannot hold.

4.3 Decisions and games

s6-nano is the s3 architecture at language scale, trained with AdamW after the ablation of section 4.9 kept none of its additions. Its pretraining read 0.87B tokens, 588 per parameter (`BENCH.md` , section "skycmd-s6"). On `typed_decisions_bench` , the test split of `LocalLLaMA/typed-decisions` (2,000 questions), it scores 0.480, above the majority class (0.398), where s3-pico (0.382) and s4-micro (0.352 on the same records) stayed below; s5-lite reaches 0.696 and oscar-1-17m 0.6775 (Results). It plays the four games at the oracle's score at full depth: Flappy 200 of 200 pipes with no crash, Rescue 100%, Dodge 87% with 23 reflex vetoes, the Doom-like game 100% of levels. It fails on what it never trained on: 0.055 on the 13-way NOTAM subject question (chance 0.077) and 0.404 on When2Call (`docs/s6-nano2-study.md` §1, §3.4).

The gate was too low, and a study made after it shows why (Results). On 17 of the bench's 20 tasks, s6-nano gives one answer to at least 95% of the items; its 0.4805 equals answering each question with its most common training answer (0.4815). The bench's tasks were 0.4% of its decision-training draws. A logistic regression on hashed character n-grams, no larger than the 715,136-byte file, reaches 0.703 on the validation split of the same tasks: a rehearsal on validation, not a test score. English MASSIVE is not strictly zero-shot (206 of its 2,000 test utterances are also in training; amendment 14 of `docs/s6-gen-prereg.md`). The verdict does not change.

s6-micro v1 (5,587,201 parameters) scored 0.547 on the bench but missed the ECE guard: 0.0821 at full depth, 0.0804 at its conformal thresholds, against 0.08. It is calibrated in distribution (ECE 0.037 on validation), and one zero-shot NOTAM statement, 412 of the 11,000 test items, weighs up to about 0.02 of its pooled test ECE (amendment 12; What we bring). **s6-micro v2**, with twice the decision-training minutes and everything else the same, passed the same gate on the same test records, 0.553 with an ECE of 0.0588; it is a second look (Results).

Before s6. s4-micro, distilled from Qwen3.5-4B in one night, scored 0.343 against a gate of 0.55 (`BENCH.md` , section "skycmd-s4"). The three pre-registered bets of s5 all missed (Table 3). The first was

the iso format of section 3.1: at equal data it was far more accurate than the usual format on the fitted bench (0.674 against 0.429, +0.245 [+0.136, +0.321], three seeds each) and less well calibrated, by more than the margin fixed in advance. The coherence losses of bet 2 improved only the identities they were trained on, 8-bit quantization could not be shown non-inferior for lack of power, and the online escalation of bet 3 failed in all six cells (Results, s5).

4.4 Secrets

s6-secrets-pico (the s3-pico graph, 250,272 bytes) reads a short text and answers four questions. On `secret_bench`, 1,996 synthetic texts in four file types the training never shows, it scores 0.9990 on "is a usable secret exposed?" and 0.9795 on "which service issued it?", against 0.977 and 0.787 for a gitleaks-style rule scanner on the same texts; its line answer is weak (0.6859, ECE 0.3051) (Results). The bench is synthetic: our generator's formats and templates, with fake keys. Measured after publication on 719 real files in which a rule scanner finds nothing, cut into 22,433 chunks, it flags 10.6% of the chunks and 59.8% of the files at p ≥ 0.9, and every flag is a false positive; its recall on real secrets was not measured (real text). The gate stays passed; the model is published as the result of its pre-registration, not as a scanner.

s6-secrets-verify-pico changes the task: a rule proposes a candidate value and the model judges it in context. Trained and tested on Samsung's CredData, on 221 test repositories its line F1 is 0.8561, against 0.7332 for the best rule scanner (Betterleaks 1.10.1), +0.1229 [0.0413, 0.3613] where the gate asked for +0.15. Six gates of seven hold, among them 31 of 8,825 clean real files flagged (0.35%) and an ECE of 0.0073 (Results). The bar of +0.15 was set with Gitleaks' published 0.442 in mind; on this test Gitleaks reads 0.7037. Rules and verifier together (s6-secrets v3), a second look on the same files, read 0.8477 and failed G1 again (Results). A diagnosis made afterwards found that most missed API lines never reach the verifier, because the candidate generator proposes nothing on them (`docs/s6-secrets4-diagnosis.md`).

4.5 Tool routing

A router reads a request and up to 7 tools, each as `name: description`, plus *none of these tools fits*. **s6-route-nano** (the s6-micro graph, chosen by its pilot rule) also reads each tool's BM25 score in its state, because amendment 1, written before training when BM25 beat the embedding baseline on validation, added BM25 to the gate. On 2,000 test slates of tools never seen in training it reaches 0.8395, above potion-base-2M with cosine (0.6675) and both trained baselines of its size, but tied with BM25 (0.8500), and it takes 38.9 ms per request in WebAssembly where 2 ms was asked. With every tool renamed `tool_<k>` it falls to chance (0.1050) (Results).

s6-route trains nothing: it multiplies, option by option, the probabilities of that file and of potion-base-2M packed to 4 bits, raised to 0.81 and renormalised, and asks the user below a top probability of 0.671047. Its fusion was first seen on s6-route-nano's already-read test, so it was pre-registered with parameters from validation and read once on fresh requests: a second look on fresh records. On 945 fresh requests about the held-out tools it reaches 0.8868, against 0.8169 for BM25 and 0.9048 for all-MiniLM-L6-v2 (91.6 MB), with an ECE of 0.0162 and 30.9 ms per request in WebAssembly; every condition held. Stated with the pass: the 60 ms latency bar was set after s6-route-nano was measured; the advantage over BM25 leans on the three largest clusters (without them +0.0202 [−0.0062, +0.0481]); with names masked it falls to 0.2497 (Results).

A diagnosis made after the gate, on validation files only, explained the masked-name collapse: in training, the right tool of every renamed slate was the only option whose description the teacher had rewritten, and *none* was never right there, so under `tool_<k>` names the model part picks the description that reads most like a rewrite. On validation, giving one distractor the rewritten description makes the model pick it on 0.8311 of the slates (`docs/s6-route4-diagnosis.md` §1). Versions 4, 5 and 6 tried to rebuild the data without such cues and each stopped at its data checks before training (Table 3). **s6-route7-nano** trained on the rebuilt data and was read once on T5, 2,000 fresh requests about the tools of 262 MCP servers from Toucan, a family no training slate holds. With the request hidden its trained part is at chance (0.1406 against 1/7), and masking the names costs the pair 0.0170; but the pair reads 0.5610 against 0.5675 for BM25, falling from 0.8802 on validation, with an ECE of 0.1845, and with the request hidden it picks the right tool on 0.1969, above its bar of 0.1929, as its untrained embedding part alone already does (0.1931) (Results). Why it falls from validation to T5 is not measured. The line is closed; s6-route stays the published router, and reads 0.5530 on T5 (reported).

4.6 End of turn in speech

s6-ear-nano tried three tasks on one audio encoder pretrained to predict Mimi codec tokens. It failed three of its four gates (end of turn 0.5681 against 0.9263 for Smart Turn v3.2), and its control without the pretraining did better on all four sets (Results).

s6-ear2-nano drops the pretraining and trains one task for 240 working minutes as a student of Smart Turn v3.2. On the 31,527 test clips of Smart Turn v3.2 it answers 0.8993, with a false cut-off of 0.0847 and an END recall of 0.8627 at its validation threshold τ_{85} = 0.7244; the teacher reads 0.9260 through the same harness, the silence timer 0.4979, and the student takes 2.39 ms per decision in WebAssembly from a 631,936-byte file (Results). Its harness was checked against Pipecat's published 0.9263 before any test score was read. Stated with the pass: the test set is partly in-sample for its design choices, and on eot-bench, LiveKit's 17,934 labelled pauses of real speech addressed to agents, it and its teacher are far weaker (AUC 0.7299 against 0.7838); at τ_{85} it answers "finished" on 0.5253 of the real ends of turn.

s6-ear3-nano keeps that graph and trains it again for 120 working minutes, half on real two-speaker English conversations of the gated otoSpeech corpus, each pause labelled from the other speaker's channel. The labels were checked before training: a second detector, Silero VAD, agreed with them on 0.9509 of the validation records (bar 0.85), and the silence after a pause did not carry the label by itself (AUC at most 0.4396, bar 0.60). On eot-bench, read 0.2 s into each pause (16,230 cuts), its AUC is 0.7197 against v2's 0.7448, −0.0252 [−0.0336, −0.0167] where at least +0.02 was asked; on 113 held-out conversation sessions (reported) it reads 0.6878 against v2's 0.6031 (Results). By the reading its pre-registration fixed before the run, the human-to-human signal was learned and does not transfer to speech addressed to agents (`docs/s6-ear3-prereg.md` §12, §17). No statement about otoSpeech goes beyond these published counts and statistics, and no record derived from it is published.

4.7 Forecasting

s6-cast-pico applies the encoder to patches of 32 points of a series and outputs nine quantiles for each of the next 64 steps. On GIFT-Eval (97 configurations, every score divided by the seasonal naive's, lower is better) it reads nGMASE 0.9471 and nWQL 0.6646, against bars of 0.838 and 0.587 taken from the training-free FLAIR; TinyCast as published reads 0.774 and 0.545 (Results). **s6-cast2-pico** adds

TinyCast's zero-parameter period detector, a seasonal draft in each decoder query, sub-hourly real series and 720 working minutes. It improves on v1 (0.8847 and 0.6190), and its MASE is under the seasonal naive's at every term (G2 holds), but it misses G1 and stays behind the project's TinyCast port (0.7918 and 0.5622 without flip); on Chronos-ZS, which its design never read (reported), it reads 0.9014 and 0.7143 against 0.8598 and 0.6928 for the port (Results). GIFT-Eval is partly in-sample for its design. Its file is 148,512 bytes against the port's 595,840, and it forecasts 64 steps in 838 µs natively against 21,531 µs, on one synthetic series and one core. The line stops; the project's forecaster is the port.

4.8 Memory

s6-sleep adds a fast-weight memory (a small matrix written once per frame and read by every later frame) and "sleep" passes that replay the last frames. Version 1 stopped at its pilot and version 2 at a CPU check (Table 3). **s6-sleep3-nano** holds the memory's gates at their initial values and adds, in training only, a loss that decodes the asked colour from the memory. On the `recall` test (a door's colour shown once, asked 1 to 32 frames later; chance 0.25) it is right on 0.4892 of the questions against 0.2440 for the same training without a memory, three seeds pooled, and the gain fades with the delay (0.6147 at 1 to 4 frames, 0.3302 at 17 to 32). Four sleep passes add +0.0030 [−0.0010, +0.0068], and the memory costs two of the four games: Flappy 15.87 against 71.05 for the control, Dodge 5.89 against 7.94 (Results). Version 4 failed both checks on training sequences and is closed (Table 3).

4.9 The LFM3-1.7B router and the ablation

System 1 / System 2. For LFM3-1.7B, a 6-coefficient probe on the answer given without thinking (System 1) decides whether to ask again with thinking on (System 2), at a threshold from conformal risk control. On 1,000 test questions of GSM8K and MATH-500, with the 1,024-token budget chosen on calibration, System 1 alone scores 0.667 at 559 tokens, System 2 alone 0.796 at 958, the router 0.786 at 1,032 and the oracle 0.814 (Results). The router keeps the accuracy (H2) but uses more tokens than always thinking (H3 fails): System 1 alone already takes more than half of System 2's tokens, and every question sent on pays for both.

The ablation. Before the big run, a pre-registered ablation at the pico size decided which of eight additions to adopt; each needed a validation gain of 0.01 (0.005 for the training options) with no game lost. None was adopted: Muon gained 0.0051 and lost Flappy score, Parallax attention 0.0054, exclusive self-attention 0.0081, and SODA weight decay, exit distillation, a weight average, StableQAT and a pretraining curriculum each lowered validation accuracy (by 0.0021 to 0.0062) (`runs/recipe/recipe.json` ; What we bring).

---

5. Observations across lines

The observations below were not pre-registered as tests. Each one is a pattern in readings already published, stated with the lines it rests on and what it does not show.

5.1 Early exits must be calibrated where the model is used

An early exit lets a looped model answer after fewer passes when it is confident. Two models give the same lesson (What we bring, section 3).

Table 4. Early-exit thresholds fitted on held-out questions against thresholds checked by playing. s2-nano and s3-pico: 5 arena seeds × 30 episodes, mean ± standard deviation over seeds (`loopcal.json` ; Doom is played by s2-nano-doom). s6-nano and s6-micro v2: 30 episodes per game (`BENCH.md` , section "skycmd-s6", Arena).

model and thresholds	Flappy pipes (of 200)	Flappy crashes (of 30)	Rescue	Dodge vetoes	mean passes
s2-nano, τ fitted on held-out questions (0.62 / 0.62 / 0.69)	15.86 ± 2.30	30 ± 0	2.98 of 3	1,931.4 ± 1,105.8	1.32 (Flappy)
s2-nano, τ = 0.95 fitted by playing	191.94 ± 7.16	2.4 ± 2.1	3.00 of 3	250.4 ± 505.7	3.43 (Flappy)
s3-pico, τ = 0.6 fitted by playing	200 ± 0	0	3.00 of 3	9.6 ± 4.1	1.44 (Flappy)
s3-pico, full depth	200 ± 0	0	3.00 of 3	9.6 ± 4.1	4
s6-nano, full depth	200	0	3.00 of 3	23	4
s6-nano, conformal thresholds (α = 0.02)	200	0	2.80 of 3	1,016	1.01 to 1.83 per game
s6-micro v2, full depth	200	0	3.00 of 3	20	6
s6-micro v2, conformal thresholds	200	0	3.00 of 3	21	2.0 to 2.7 of 6

- **Evidence.** s2-nano, at thresholds fitted on held-out questions drawn independently of play, crashed in every Flappy episode on all five seeds; the same weights at a threshold chosen by playing flew 191.94 pipes on average. s6-nano's thresholds came from conformal risk control, which bounds at α = 0.02 how often an early answer differs from the full-depth answer on held-out questions; played at those thresholds, Rescue fell to 93% (2.80 of 3) and the Dodge vetoes rose from 23 to 1,016, so the pre-registered rule rejected them and the file ships at full depth. s6-micro v2's conformal thresholds were rejected too, by one extra veto (Results).
- **Why it can happen.** Playing with early exits changes which states the model visits, so a guarantee fitted on questions drawn from one distribution does not cover the states its own decisions produce.
- **Limits.** The evidence is two models where the fitted thresholds failed in closed loop. For s3-pico the check confirmed that its exits were safe rather than rescuing anything, and s6-micro v2 lost no score at its conformal thresholds. The s2-nano comparison is an exploratory run, not a pre-registered test; s6-nano's closed-loop check was a pre-registered rule. All of it is on four friendly games whose useful features are computed by code.

5.2 Readings on data built differently from training limited the passes; most failures came on the line's own test

Where one file was read both on its own distribution and on data built differently, the second reading is often far lower. Table 5 gathers the pairs that the published pages and result files hold. It is assembled after the fact from published numbers and is not a pre-registered test; the metrics differ from row to row, and so does what "other data" means.

Table 5. The same file read on its own distribution and on other data. Differences computed here are marked *c*.

line	own distribution	other data	difference
s6-secrets-pico (passed)	synthetic bench: secret 0.9990	719 real files with no secret: 59.8% flagged at p ≥ 0.9	(different metrics)
s6-route (passed)	fresh requests, held-out tools: 0.8868	T5, a new tool family (reported): 0.5530; BM25 0.5675	−0.3338 *c*
s6-ear2-nano (passed)	Smart Turn's test clips: AUC 0.9600	eot-bench: AUC 0.7299; teacher 0.7838	−0.2301 *c*
s6-nano (passed)	`typed_decisions_bench`: 0.480	NOTAM subject, no training data: 0.055 (chance 0.077)	(different tasks)
s6-route7-nano (failed)	validation: 0.8802	T5: 0.5610; BM25 0.5675	−0.3192 *c*
s6-ear3-nano (failed)	held-out conversations: AUC 0.6878, v2 0.6031	eot-bench at +0.2 s: AUC 0.7197, v2 0.7448	against v2: +0.0847 *c* and −0.0252
s6-micro-v1 (failed)	validation: ECE 0.037	pooled test with zero-shot NOTAM: ECE 0.0821	+0.0451 *c* (lower is better)
s6-secrets-verify-pico (failed)	validation repositories: line F1 0.8603	221 test repositories: 0.8561	−0.0042 *c*
s6-route-nano (failed)	validation slates, after the gate: 0.8339	test slates of unseen tools: 0.8395	+0.0056 *c*

Sources beyond Results: s6-secrets-verify-pico's validation from `BENCH.md` (section "s6-secrets-verify-pico"); s6-route-nano's validation from `docs/s6-route4-diagnosis.md` §1; s6-micro-v1's validation ECE from amendment 12.

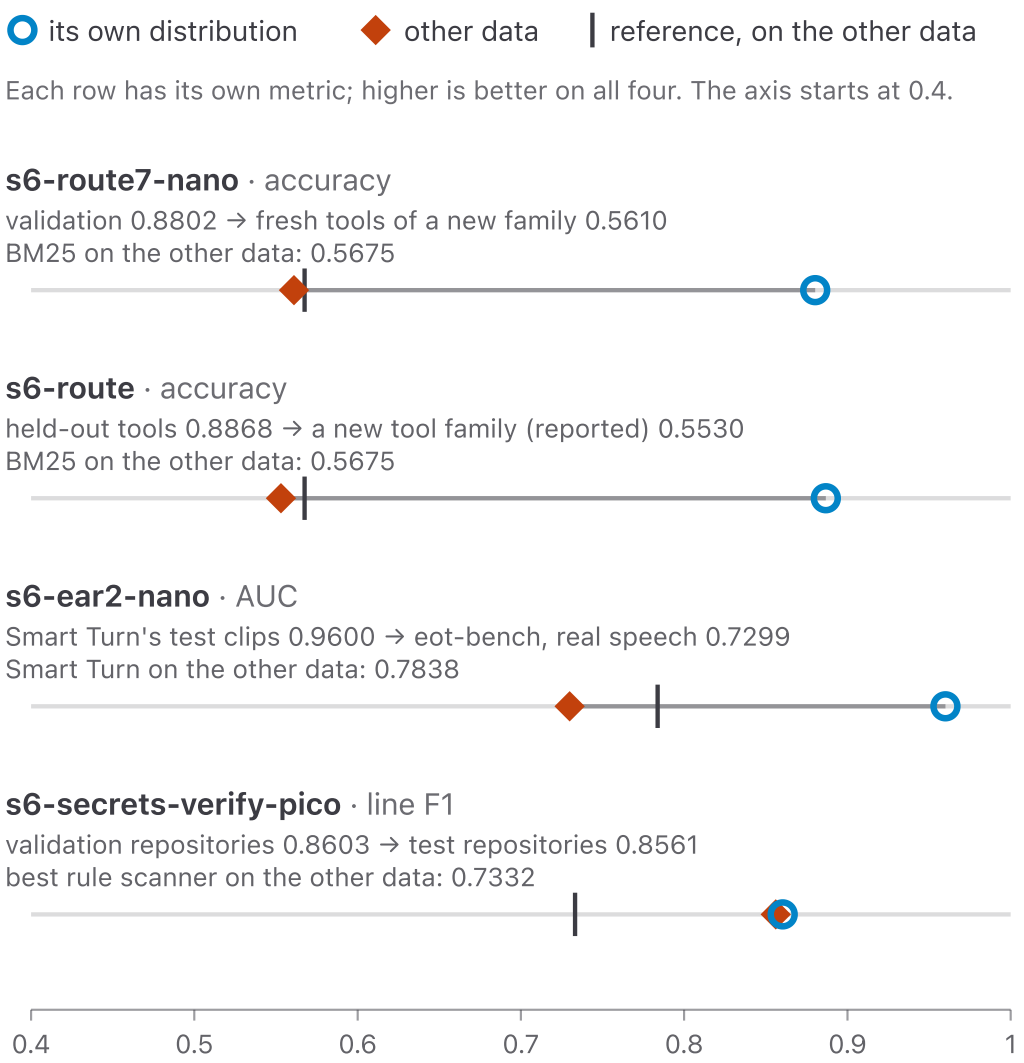


Figure 1. Four rows of Table 5 whose two readings share a metric: the line's own reading (ring), the reading on other data (diamond) and a reference on that other data (tick). Each row has its own metric, higher is better on all four, and the axis starts at 0.4. The figure is written by `scripts/paper-figures.py` from the numbers of Table 5; its text alternative is in the SVG.

- **What it shows.** Three of the five passes carry a limit measured on other data (s6-secrets-pico, s6-ear2-nano, s6-route), and s6-nano fails where it had no training data. Two failures were decided by a reading on other data: s6-route7-nano on a new tool family, and s6-ear3-nano on speech addressed to agents after it had improved on its own corpus. s6-micro-v1's calibration failure traces largely to a zero-shot source.
- **What it does not show.** Most failures did not come from transfer. s6-ear-nano, s6-sleep3-nano, s6-cast-pico, s6-cast2-pico, s4-micro, the LFM3 router and the s5 bets missed on the test the line was built for; s6-secrets-verify-pico and s6-route-nano lost nothing from validation to test and failed against their baselines. The forecasters were tested only on series they never trained on, so they have no in-distribution reading to compare. The causes of the falls are not measured: for s6-route7-nano, no reading separates the new tool family from the new source of requests or from T5's construction.

5.3 A gate must be checked against trivial baselines before it is frozen

- **s6-nano.** The bar was the majority class over the whole bench (0.398). Answering each question with its most common training answer, a rule that reads no state, scores 0.4815, above s6-nano's 0.4805; a size-matched linear model reaches 0.703 on the validation split of the same tasks. Both were measured after the gate (Results). A gate set against the per-question prior would have failed.
- **s6-route-nano.** Here the check happened before training: amendment 1 added BM25 to the gate when BM25 beat the planned embedding baseline on validation. The gate then caught a router that ties a lexical score: −0.0105 [−0.0265, +0.0055] against BM25, where the planned bar on the embedding baseline, potion-base-2M + 0.03, was cleared (0.8395 against 0.6975) (Results). BM25 stayed above the retrained router on the new tool family (section 4.5).
- **s6-secrets-verify-pico.** The margin of +0.15 was set with Gitleaks' published 0.442 in mind; on the test, Gitleaks reads 0.7037 (Results). A bar set from a published number, rather than rehearsed on the line's own validation data, ended up harder than intended.
- **Limits.** Three lines. In one the trivial baseline was found too late, in one it was found in time, and in one the bar moved the other way. Other gates did include a trivial baseline from the start (the silence timer for end of turn, 0.4979; the seasonal naive for forecasting).

5.4 Controls that hide or rename the input expose shortcuts

- **Renamed options.** Masking tool names as `tool_<k>` took s6-route-nano from 0.8395 to 0.1050 and s6-route from 0.8868 to 0.2497, where BM25 kept 0.8445 and 0.8138; the diagnosis traced it to the teacher's rewritten description marking the right tool in every renamed training slate (section 4.5). s6-route7-nano, trained without that cue, lost only 0.0170 under masking (Results).
- **The request hidden.** With the request replaced by `?`, a router that reads the request should be at chance. The published s6-route files read 0.1104 to 0.1402 on validation and fresh slates, at or below a chance of 0.1429 (Results); s6-route7-nano's trained part read 0.1406, but its pair read 0.1969, above the bar of 0.1929, because the untrained embedding already did (0.1931).
- **Rules that never read the request.** Counts made with no model found that the right tool's description is the slate's shortest far more often than chance: a rule picking it is right on 0.2439 of s6-route-nano's test slates whose answer is a tool and on 0.3633 of s6-route's fresh ones, against 1/7; whether the models used it is not measured (study). Pre-registered data checks of the same kind stopped versions 4, 5 and 6 before any training (Table 3).
- **Controls that passed.** For s6-ear3-nano, the label cross-check (0.9509) and the probe showing the silence did not carry the label (AUC at most 0.4396) held before training, and the version still failed on its gate (section 4.6).
- **Limits.** All the shortcut findings come from one line, tool routing, and most were found after its gates. The rules count what a cue could buy, not what a model used.

5.5 What the negative results are

Three results are negative in the sense that an addition did not do what it was built for.

- **The ablation** adopted none of its eight additions at the pico size (section 4.9). This says that none brought 0.01 of validation accuracy there, under these budgets; it does not say they cannot help elsewhere.
- **The memory is used, and sleep adds nothing measurable.** s6-sleep3-nano's memory gains +0.2452 on `recall`, its sleep passes +0.0030 with an interval holding zero, and the memory costs two games (section 4.8). One probe, one size, three seeds.
- **The System 1 / System 2 router keeps accuracy but not the token budget**, because the answer without thinking is not cheap enough for routing to pay (section 4.9). One model, maths questions only.

---

6. Limitations

- **One training run per gated model.** Except s5 (three seeds per format) and s6-sleep3-nano (three training seeds, pooled), every gated model was trained once, on one seed. The one line read on two training seeds, s6-sleep v4, met its last bar on one seed of the two (Results).
- **Intervals are over test items, not over training runs.** They say how much a number would move on other test items, not how much it would move if the same recipe were trained again.
- **Own benchmarks and friendly games.** s6-secrets-pico's bench is the project's own generator. The games compute the useful features (relative clearances, an A* waypoint) and leave the model to decide, and the useful baselines there were not run: a same-size MLP on the same features, an RL policy, oscar-1-17m on the games (What we bring).
- **Second looks and in-sample designs.** s6-micro v2 and s6-secrets v3 were judged on test records already read; s6-route was designed on a test already read; Smart Turn's test is partly in-sample for s6-ear2-nano and s6-ear3-nano, and GIFT-Eval for s6-cast2-pico.
- **Gates that were too low or too loose.** s6-nano's gate is passed by a constant answer per task (section 5.3); s6-route's latency bar was set after its predecessor was measured.
- **Baselines not run.** For s5, the baselines B1 (label log-probability through the MLM head) and B3 (a linear head on the frozen backbone) were not run, a deviation declared before any result (Results). Ablations of s3 over several seeds and a standard control benchmark with vector states were not done.
- **The literature search was not systematic**: two research passes.
- **Earlier models have no gate.** s1-pico, s2-nano, s2-nano-doom and s3-pico were built before the protocol; their numbers are measurements, not gated results.
- **One project, one machine, no outside review.** The protocol was written and applied by the same project: nobody independent froze the pre-registrations, opened the tests or checked the amendments, and some test-opening markers were created under the owner's standing delegation rather than by the owner's hand, as the result blocks state. Nothing here has been peer-reviewed.
- **The observations of section 5 are post hoc.** They were assembled after the readings and are not tests.

7. Reproducibility and artefacts

- **Code and documents.** The repository is github.com/maxgfr/skycmd, Apache-2.0 for the code. The pre-registrations, with their amendments and result blocks, are in `docs/`, indexed in `docs/README.md`; every measured number is in `BENCH.md`; the small result files of each gated run are under `runs/`.
- **Model files.** Every published file is committed under `site/public/models/` and listed in the manifest with its size and verdict. Every manifest entry also records the sha256 of its file, or of each of its two files, and the site's tests check each one against the committed file (`site/tests/manifest.test.ts`); several result blocks record the same hash for the file they published (for example amendment 17 of `docs/s6-gen-prereg.md` for s6-micro and §17 of `docs/s6-ear3-prereg.md` for s6-ear3-nano). Each file runs in the browser on this site's demo pages.
- **Hugging Face.** Each of the 20 published models has a model repository `maxgfr/<name>` whose card opens with its verdict, and a Space of the same name; the Space `maxgfr/skycmd` gathers every demo (`hf/spaces.json`). The 16 data sets are listed on Sources. Since 20 November 2026, `maxgfr/skycmd-decisions` holds both splits of `LocalLLaMA/typed-decisions` (`typed_decisions_train`, and `typed_decisions_bench` for evaluation only), and `maxgfr/skycmd-glm-labels` all 747,077 GLM-5.3-flash labels, marked as generated by a model.
- **What is not redistributed, and why.** Records derived from the gated otoSpeech corpus, whose access is granted per user; sets of real credentials (CredData's values are obfuscated upstream, and its files keep their own projects' licences); the s6-ear audio features (31 GB, rebuilt from their public sources); and evaluation sets used as given, such as GIFT-Eval (Sources, `DATA_LICENSES.md`).
- **Commands.** The README's table "Reproduce a result" gives, for each s6 model, its script and result files (`README.md`); for example `scripts/train-turn.sh` for s6-ear2-nano or `scripts/route3-read.sh` for s6-route, which trains nothing. The scripts read their data from a folder next to the repository, described in `docs/data-inventory.md`. The figure of this report is rebuilt by `python3 -I scripts/paper-figures.py`.
- **Hardware.** One Apple M5 laptop with 16 GB of memory, under the brake of section 3.2; latencies were measured on the same machine, often while another job ran, as each result says.

References

1. mgoeckel. *oscar-1-17m*, a decision model on Ettin-17M. Hugging Face model card. huggingface.co/mgoeckel/oscar-1-17m
2. Convai Innovations. *Laya*. Hugging Face model card. huggingface.co/convaiinnovations/laya
3. Supersonic Labs. *Julia-1*. Hugging Face model card. huggingface.co/SupersonicLabs/Julia-1
4. Y. Sun, J. Xu, J. Shi, Z. Yang. *Type-Safe Is Not Error-Free: Typed Decision Models Follow the Option Name, Not the Definition Bound to It.* 2026. arXiv:2609.26758
5. G. Amit. *Permutation-Robust Decision Modeling with Candidate-Independent Block-Causal Attention.* 2026. arXiv:2610.01601

6. A. Jolicoeur-Martineau. *Less is More: Recursive Reasoning with Tiny Networks.* 2025. arXiv:2510.04871
7. R.-J. Zhu, Z. Wang, K. Hua, et al. *Scaling Latent Reasoning via Looped Language Models.* 2025. arXiv:2510.25741
8. Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko. *Revisiting Deep Learning Models for Tabular Data.* 2021. arXiv:2106.11959
9. T. Zhou, D. Fu, M. Soltanolkotabi, R. Jia, V. Sharan. *FoNE: Precise Single-Token Number Embeddings via Fourier Features.* 2025. arXiv:2502.09741
10. S. Ma, H. Wang, L. Ma, et al. *The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits.* 2024. arXiv:2402.17764
11. J. Nielsen, L. Galke, P. Schneider-Kamp. *When are 1.58 bits enough? A Bottom-up Exploration of BitNet Quantization.* 2024. arXiv:2411.05882
12. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger. *On Calibration of Modern Neural Networks.* ICML 2017. arXiv:1706.04599
13. A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, T. Schuster. *Conformal Risk Control.* 2022. arXiv:2208.02814
14. T. Schuster, A. Fisch, J. Gupta, M. Dehghani, D. Bahri, V. Q. Tran, Y. Tay, D. Metzler. *Confident Adaptive Language Modeling.* NeurIPS 2022. arXiv:2207.07061
15. M. Jazbec, A. Timans, T. Hadži Veljković, K. Sakmann, D. Zhang, C. A. Naesseth, E. Nalisnick. *Fast yet Safe: Early-Exiting with Risk Control.* 2024. arXiv:2405.20915
16. A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, L. Lei. *Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control.* 2021. arXiv:2110.01052
17. S. Ross, G. J. Gordon, J. A. Bagnell. *A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning.* AISTATS 2011. arXiv:1011.0686
18. K. Wen, D. Hall, T. Ma, P. Liang. *Fantastic Pretraining Optimizers and Where to Find Them.* 2025. arXiv:2509.02046
19. M. Akhtar, H. N. Nounou, M. Mansouri, M. N. Nounou. *When Does Muon Actually Win? A Paired Data-, Time-, and Tuning-Budget Benchmark.* 2026. arXiv:2610.21555
20. Pipecat. *Smart Turn v3.2*, an open end-of-turn model (BSD-2-Clause). github.com/pipecat-ai/smart-turn
21. LiveKit. *eot-bench*, an end-of-turn benchmark in 14 languages (CC BY 4.0). huggingface.co/datasets/livekit/eot-bench-data
22. T. Aksu, G. Woo, J. Liu, X. Liu, C. Liu, S. Savarese, C. Xiong, D. Sahoo. *GIFT-Eval: A Benchmark For General Time Series Forecasting Model Evaluation.* 2024. Data: huggingface.co/datasets/Salesforce/GiftEval
23. A. F. Ansari, L. Stella, C. Turkmen, et al. *Chronos: Learning the Language of Time Series.* 2024. arXiv:2403.07815; data: huggingface.co/datasets/autogluon/chronos_datasets
24. RAWS Labs. *TinyCast*, a 146,505-parameter forecaster (Apache-2.0). Hugging Face model card. huggingface.co/raws-labs/tinycast

25. Samsung. *CredData*, a credential data set of open-source repositories. github.com/Samsung/CredData
26. Minish Lab. *potion-base-2M*, a Model2Vec static embedding (MIT). huggingface.co/minishlab/potion-base-2M
27. sentence-transformers. *all-MiniLM-L6-v2*. Hugging Face model card. huggingface.co/sentence-transformers/all-MiniLM-L6-v2
28. Agent-Ark. *Toucan-1.5M*, tool-agent trajectories over real MCP servers (Apache-2.0). huggingface.co/datasets/Agent-Ark/Toucan-1.5M
29. NVIDIA. *When2Call* (CC BY 4.0). huggingface.co/datasets/nvidia/When2Call
30. S. Lee, S. McLeish, T. Goldstein, G. Fanti. *Do Language Models Need Sleep? Offline Recurrence for Improved Online Inference.* 2026. arXiv:2605.26099

How to cite

bibtex

```bibtex
@misc{maxgfr2026skycmd,
  author       = {maxgfr},
  title        = {Pre-registered gates for small models trained on one laptop:
                  what passed, what failed, and what the readings show},
  year         = {2026},
  howpublished = {\url{https://maxgfr.github.io/skycmd/paper/}},
  note         = {Technical report, not peer-reviewed}
}
```

The same citation is in `CITATION.cff`. The report is updated when a result changes; the date at the top is that of its last change.