

Experiment Receipt Discipline for Small-Model Autoresearch

Status: DRAFT **Date seeded:** 2026-07-06 **Author:** Reuben Bowlby (ORCID 0009-0002-5620-1103)
Repository: hummbl-dev/hummbl-papers **Source packet:** hummbl-dev/research-source-packets#9 (candidate)

Abstract

Small-model autoresearch loops can generate dozens of experiments per session, but without receipt discipline the results become unauditable: hypotheses blur into outcomes, baselines drift, noise floors go unmeasured, and negative results disappear. This paper presents a case study of the `autoresearch-pipeline` corpus — 243 result rows, 55 kept patches, 186 DISCARD negative results, 2 crashes — as a worked example of receipt-bearing autoresearch. The strongest finding is not a new architecture patch; it is the governance value of a *prospective* noise-floor gate that blocks architecture experiments before the noise floor is measured, and the preservation of negative results as first-class evidence. We present the QK-Norm noise-floor-gate vignette as the centerpiece: a case where the noise floor (0.180 untuned) was 36–90× larger than the expected improvement (0.002–0.005 bpb), which the receipt discipline caught *before* any architecture experiment was authorized.

1. Problem

Autoresearch loops at small scale produce many experiments with small effect sizes. Without receipt discipline, the loop becomes vulnerable to four failure modes:

1. **Cherry-picking:** keeping only the improvements and silently dropping the failures, creating a survivorship-biased record.
2. **Baseline drift:** the reference point shifts across experiments without tracking, making comparisons meaningless.
3. **Noise-floor blindness:** running architecture experiments without measuring run-to-run variance, so improvements within the noise floor are indistinguishable from real effects.
4. **Frontier-feature cargo-culting:** importing techniques from large-model regimes (e.g., QK-Norm, value residuals) where the dynamics differ, without testing whether they transfer to the small-scale regime.

These failure modes are well-known in the ML reproducibility literature (Pineau et al. 2021; Sculley et al. 2018). The contribution of this paper is not the invention of receipts or reproducibility checklists; it is the application of receipt discipline to a *governed agent loop* at small scale, with a *prospective* noise-floor gate that prevents architecture experiments from being authorized before the noise floor is measured.

2. Prior Art

The receipt-discipline approach builds on three traditions:

- **ML reproducibility checklists** (Pineau et al. 2021): the NeurIPS Reproducibility Program established retrospective documentation standards for ML experiments. Our receipts are prospective (gated before execution) rather than retrospective (documented after the fact).
- **Pre-registration standards** (OSF; PCI RR): the registered-reports model pre-declares hypotheses and analysis plans before data collection. Our noise-floor gate is analogous — it pre-declares the variance threshold before architecture experiments are authorized.
- **Notebook conventions** (Pitt-Wipf): the observation/interpretation separation keeps raw observations distinct from interpretive commentary. Our receipt schema enforces this by separating the measured outcome (observation) from the KEEP/DISCARD decision (interpretation).

The contribution is the *application* of these principles to a governed agent loop, not the principles themselves. A human experimenter can be held accountable post-hoc; an autonomous agent loop needs the discipline encoded as a gate, or it will silently drop failures.

3. The Receipt Schema

Each governed experiment records seven fields:

Field	Description	Example
Hypothesis	What the experiment tests	“QK-Norm improves val_bpb at 26M scale”
Baseline	The reference result, with full config	val_bpb=0.459677, may3 config (TO-TAL_BATCH_SIZE=2 ¹⁴ , 1014 steps)
Noise floor	Run-to-run variance under control conditions, with N and methodology	Range=0.004988, N=3 seeds, std=0.001843, CV=0.38%
Measured outcome	The val_bpb, step count, and config the experiment produced	val_bpb=0.484, 995 steps, seed=42
Decision	KEEP or DISCARD, with a bounded reason	DISCARD: within noise floor (delta < 0.003)
Config snapshot	Full hyperparameters (LR, batch size, seed, architecture)	LR=0.04, batch=2 ¹⁴ , seed=42, depth=6
Provenance	Git commit hash, hardware, timestamp	commit=8deb013, RTX 4000, 2026-07-05

The first five fields were present in the original schema; the config snapshot and provenance fields are additions prompted by the Pineau checklist, which requires code/data provenance and full hyperparameter documentation. A receipt without config and provenance is not reproducible.

Receipt Contract

- One receipt per governed experiment
- Seven required fields (hypothesis, baseline, noise floor, outcome, decision, config, provenance)
- Negative results preserved, not dropped
- Noise floor measured *before* authorizing architecture experiments (prospective gate)
- Scale-dependence annotated per experiment idea pack
- Decision vocabulary is controlled: KEEP (improvement above noise floor), DISCARD (within noise floor or no improvement), CRASH (execution failure)

4. The QK-Norm Noise-Floor Gate (Centerpiece)

The strongest example of receipt discipline preventing a failure mode is the QK-Norm noise-floor gate.

Context

The autoresearch-pipeline supervisor proposed testing QK-Norm (normalizing Q and K projections before attention) at 26M parameter scale. The expected improvement from the ML literature was -0.002 to -0.005 bpb (bits-per-byte).

What the receipt discipline caught

Before authorizing the QK-Norm experiment, the supervisor ran two seed-repeat control experiments to measure the noise floor:

Experiment 0 (untuned config): Master HEAD config (TOTAL_BATCH_SIZE=2¹⁹, 42–44 steps), 5 seeds.

Statistic	Value
Mean val_bpb	0.921964
Std dev	0.080456
Range (noise floor)	0.180419
CV	8.73%

Verdict: FAIL. The noise floor (0.180) was 36–90× larger than the expected QK-Norm improvement (0.002–0.005). Running the QK-Norm experiment against this config would have produced an uninterpretable result — any observed delta would be within the noise.

Experiment 0b (tuned config): May3 config (TOTAL_BATCH_SIZE=2¹⁴, ~2100 steps), 3 seeds.

Statistic	Value
Mean val_bpb	0.480367
Std dev	0.001843
Range (noise floor)	0.004988
CV	0.38%

Verdict: MARGINAL. The noise floor (0.005) was at the upper boundary of the expected QK-Norm improvement (0.002–0.005). QK-Norm was authorized *conditionally*: run with ≥ 3 seeds, use 2-sample t-test ($p < 0.05$), only claim improvement if mean delta > 0.003 .

Why this matters

Without the noise-floor gate, the QK-Norm experiment would have been authorized against the untuned config, where the noise floor was 36–90× the expected effect. The result would have been a random number within [0.74, 0.92] — uninterpretable, but potentially reported as “QK-Norm doesn’t help” or worse, “QK-Norm hurts” if the random draw happened to be above the mean. The receipt discipline caught this *before* the experiment ran.

Limitations of this measurement

The noise floor estimates are based on N=5 (untuned) and N=3 (tuned) seeds. These are point estimates, not confirmed variance distributions. A proper noise floor characterization would use ≥ 10 seeds and report the full distribution. However, even with N=3, the 36× gap between the untuned noise floor and the expected effect is large enough to be confident that the untuned config is unsuitable — the gate’s decision (block untuned, conditionally allow tuned) is robust to the small sample size.

5. Corpus Quantification

The autoresearch-pipeline corpus at commit f39a3ee contains:

Metric	Value
Total result rows in <code>results.tsv</code>	243
Governed experiments	34
Kept patches (KEEP)	55
Discarded results (DISCARD)	186
Crashes	2
Noise floor measurements	2 (untuned + tuned)
Receipt documents	9 (per-wave)
Pack 2 frontier port experiments	5 (all DISCARD)

Discard categorization

The 186 DISCARD results break down as follows:

Category	Count	Description
Hyperparameter sweep discards	~120	LR, batch size, warmdown, depth variations that didn't improve
Architecture discards	~30	SwiGLU, GQA, window pattern, MLP expansion — no improvement
Pack 2 frontier port discards	15	QK-scale, resid-decay, x0-decay, smear, backout, GELU (3 seeds each $\times$ 5 features)
Seed-repeat controls	~6	Noise floor measurement runs (not experiments, but recorded)
Other	~15	Various one-off tests

The 55 KEEP results are improvements within the small-scale regime (batch size, LR tuning, warmdown ratio, depth reduction). They are not generalizable architecture claims.

Receipt completeness audit

We audited the 34 governed experiments against the 7-field receipt schema:

Field	Completeness
Hypothesis	34/34 (100%)
Baseline	34/34 (100%)
Noise floor	2/34 (6%) — only the seed-repeat controls measured it; subsequent experiments inherited the tuned noise floor
Measured outcome	34/34 (100%)
Decision	34/34 (100%)
Config snapshot	24/34 (71%) — early experiments lacked full config; later experiments included it
Provenance	34/34 (100%) — all experiments have commit hashes

The noise-floor field is the weakest: only 2 of 34 experiments directly measured it. The remaining 32 inherited the tuned noise floor (0.005) from Experiment 0b. This is acceptable for experiments within the same config regime, but the inheritance should be explicit in the receipt, not implicit.

6. The Pack 2 Frontier Port (Scale-Dependence Case)

The Pack 2 nanochat frontier port tested 5 frontier features at 26M scale, each with 3 seeds (42, 137, 271):

Feature	Seed 42	Seed 137	Seed 271	Decision
GELU activation	0.4918	0.4872	0.4845	DISCARD
Smear (learned bigram mixing)	0.4845	0.4852	0.4886	DISCARD
x0-decay (lambda decay)	0.4813	0.4809	0.4826	DISCARD
QK-scale (factor 1.2)	0.4820	0.4831	0.4817	DISCARD
Resid-decay (lambda decay)	0.4801	0.4805	0.4803	DISCARD

Feature	Seed 42	Seed 137	Seed 271	Decision
Backout (mid-layer residual subtraction)	0.4874	0.4853	0.4839	DISCARD

All 5 features were DISCARD. The best result (resid-decay, seed 42: 0.4801) was within the noise floor (0.005) of the baseline (0.4804). This provides preliminary evidence that some frontier features may not transfer to 26M scale — a scale-dependence lesson preserved by the receipt discipline rather than silently lost.

Important caveat: This is a preliminary observation from a single implementation at a single scale (26M parameters, ~2100 steps, 5-minute budget), not a generalizable claim about frontier feature transfer. The scaling laws literature (Kaplan et al. 2020; Hoffmann et al. 2022) establishes that model behavior is scale-dependent; our observation is consistent with that but does not extend it. What scale-dependence thresholds separate “frontier feature applies” from “frontier feature is cargo-culting” remains an open question.

7. Limitations

1. **Single-operator, single-corpus.** All experiments were designed, executed, and adjudicated by the same operator. There is no independent verification. An external replication protocol is provided in `replication_protocol.md`.
2. **Noise floor estimated from N=3–5 seeds.** The tuned noise floor (0.005) is based on 3 seeds; the untuned (0.180) on 5. These are point estimates, not confirmed variance distributions. A proper characterization would use ≥ 10 seeds.
3. **Scale-limited.** The corpus covers only the 26M / short-training regime. We do not assert that findings generalize to other scales. The Pack 2 frontier port results are consistent with scale-dependence but do not establish a threshold.
4. **Receipt completeness is imperfect.** The noise-floor field is inherited rather than directly measured for 32/34 experiments. Config snapshots are missing for 10/34 early experiments.
5. **The 55 KEEP results are not generalizable.** They are improvements within the small-scale regime (batch size, LR, warmdown). They should not be cited as architecture contributions.

8. Evaluation Plan

The receipt-completeness audit (§5) is the first empirical pass. Future work should test whether receipt-bearing autoresearch reduces:

- Silent dropping of negative results (currently: 0% — all 186 discards are preserved in `results.tsv`)
- Unmeasured noise floors before architecture experiments (currently: 0% — the gate blocked QK-Norm)
- Frontier-feature cargo-culting at scales where it does not apply (currently: the Pack 2 discards preserve the negative result)
- Disagreement between local status summaries and the receipt corpus (not yet measured)

9. External Replication

An external replication protocol is provided in `replication_protocol.md`. The protocol specifies the minimum hardware (consumer RTX GPU), the exact config, the seed-repeat procedure, and the pre-registered predictions. The `autoresearch-pipeline` repository will be made public for replication. An independent replicator with no HUMMBL affiliation can re-run the noise-floor measurements and the Pack 2 frontier port, then compare their results to the pre-registered predictions.

Public Source Artifacts

- autoresearch-pipeline repository (currently private; will be made public for replication)
- `program.md` — experiment program, baseline receipt, noise floor measurements
- `results.tsv` — 243 result rows, 34 governed experiments
- `docs/experiment_receipts/` — 9 per-wave receipt docs
- `research-source-packets#9` — candidate source packet
- Research notebook
- Replication protocol

Open Questions

- What is the smallest public replay fixture that demonstrates receipt-discipline value without exposing private runtime?
- What scale-dependence thresholds separate “frontier feature applies” from “frontier feature is cargo-culting”?
- Can the receipt schema be validated automatically (JSON Schema), or does it require human review?
- Should the noise-floor gate be a hard block (no architecture experiments until measured) or a soft warning?
- How many seeds are needed to characterize a noise floor reliably at this scale?

References

- Pineau, J. et al. (2021). Improving Reproducibility in Machine Learning Research. JMLR 22(164).
- Sculley, D. et al. (2018). Winner’s Curse? On Pace, Progress, and Empirical Rigor. ICLR Workshop on Reproducibility.
- Kaplan, J. et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361.
- Hoffmann, J. et al. (2022). Training Compute-Optimal Large Language Models. arXiv:2203.15556.
- Pitt, D. & Wipf, D. (2020). Observation/Interpretation Separation in Research Notebooks.
- Center for Open Science. Open Science Framework: Pre-Registration Standards.
- Peer Community In. PCI Registered Reports: Guide for Authors.