From Seed Lotteries to a Shippable Local Tab Namer: A Staged Study of Data, Decoding, and Evaluation

MD Ishtiaque Hossain
PorkiCoder Research, Vancouver, Canada

18 August 2026

Abstract

Naming a coding-agent terminal tab is a small generation task with hard product constraints: the title must be locally generated, useful, short enough to scan, and cheap enough to run on every new interaction. This paper consolidates a ten-study development campaign that began with a 12.7M-parameter T5 student and ended with a 76.96M-parameter task-tuned FLAN model selected for shipment. The campaign separates model weights from candidate construction, decoding, deterministic post-processing, data selection, and evaluation. An initial checkpoint appeared to reduce repeated-token titles by 37%, but a deterministic 31-run matched control did not support the proposed literary sequence-order mechanism. On a Stack Overflow board, a deterministic centroid highlighter and constrained four-beam search turned an unchanged 35M model into a strong relative system: on a sealed 1,000-row comparison it scored 6.20 versus 5.45 for title-tuned FLAN with the same highlighter, a paired difference of +0.752. However, an independent Codex audit rated only 19 of 40 outputs useful, below a preregistered 60% absolute gate. The same 35M stack then trailed FLAN by 1.17 points on a separate terminal-task board. Selective judged supervision produced small replicated gains, while 175,642 legal synthetic labels reduced cross-entropy to 0.91 without improving judged utility. A 24.5M pointer system reached 3.88 ms per title and beat the 35M decoder by +0.14 on one packet, but a task-tuned 77M FLAN model beat it by +0.50. Finally, leak-checked continuation training improved the frozen prior ship from 7.38 to 7.59 and from 7.34 to 7.55 on two never-seen 1,000-row holdouts. The central finding is that narrow local generation is a system-design problem: data selectivity, domain, decoding, glue, capacity, and evaluation protocol can each change the conclusion, and their effects are strongly competence-dependent.

1 Introduction

Modern coding assistants often create many concurrent sessions. A user may have one terminal agent fixing deployment, another reviewing a pull request, and a third investigating an error. Generic labels such as "Terminal 7" or unstable excerpts from the last message make those sessions difficult to recover. PorkiCoder therefore needs a local model that maps a task description to a concise tab name such as *Review Kimi Changes*, *Fix SSE Disconnect*, or *Explain DNS*. The desired output is only one to three words, but the operating requirements are unusually strict: low latency, bounded length, privacy-preserving local inference, stable formatting, and semantic usefulness without a human reference title at inference time.

This apparently small problem exposed several recurring difficulties in applied language-model research. A single favorable checkpoint suggested a mechanism that disappeared under matched

controls. A deterministic highlighter sometimes contributed more utility than the neural generator. A decoding change reversed the ranking of frozen checkpoints. A system that beat a larger benchmark on Stack Overflow pages failed an absolute usefulness gate and transferred poorly to real terminal tasks. More synthetic labels lowered cross-entropy but not utility. A compact structured head improved the quality-latency frontier but did not close the capacity gap. Finally, a larger task-tuned model won only after the training mix, leakage controls, and evaluation protocol were tightened.

The studies were published sequentially as detailed research notes between 12 and 17 August 2026 [17–26]. This paper turns that sequence into one empirical account. It does not present the campaign as a single preregistered experiment. Instead, it distinguishes exploratory findings, same-packet matched comparisons, sealed tests, and replicated fresh-holdout results.

The paper makes four contributions:

1. It provides a layer-by-layer decomposition of a local short-form generation system into weights, candidate construction, decoding, deterministic glue, data, and evaluation, then uses the decomposition to interpret a ten-study intervention sequence.
2. It documents negative results that materially changed the project: a failed seed mechanism, failed continuation training after a decoder improvement, a failed absolute usefulness gate, failed domain transfer, and a 175,642-example synthetic-data run that improved likelihood but not judged utility.
3. It quantifies the quality-latency-capacity trade-off among a 35M encoder-decoder, a 24.5M pointer system, lightweight pickers, and a 76.96M task-tuned FLAN model under a common product packet.
4. It reports the campaign's strongest result on two never-seen 1,000-row terminal holdouts, where the final unchanged-architecture continuation improved the previous ship by +0.21 and +0.20 mean judge points.

The practical conclusion is not that a particular architecture always wins. It is that the dominant intervention changes with model competence and domain. Constrained decoding and deterministic glue were decisive for a weak 35M generator; the same glue became harmful for a stronger free-form model on fresh terminal tasks. Selective labels were more valuable than a much larger weakly filtered corpus. Once the task dialect was learned, additional capacity became worth its latency and payload cost.

2 Related Work

Text-to-text transfer and instruction tuning. The models in this campaign inherit the encoder-decoder formulation of T5 [1]. FLAN showed that instruction tuning can improve zero-shot task following [2], while later work studied the scaling and composition of instruction-finetuning mixtures [3]. Our result is complementary: stock FLAN was poor at the tab-naming dialect, while a small amount of task-aligned supervision transformed it. General instruction-following ability was useful initialization, not a complete product solution.

Distillation and synthetic supervision. Knowledge distillation transfers behavior from a larger teacher to a smaller student [6]. Synthetic instruction generation can substantially expand task coverage [14], but label quantity is not equivalent to utility. Our campaign includes both selective judged distillation and a much larger teacher-label experiment, allowing a direct applied comparison between filtering strength and corpus scale.

Gap-sentence generation and extractive structure. The 35M model used gap-sentence-generation-style pretraining inspired by PEGASUS [4]. Later, a pointer head replaced autoregressive decoding with scoring over a legal title schema, following the broad idea of pointer networks [5]. The task is unusually compatible with extractive and semi-extractive structure because useful titles often reuse task words, while a small intent vocabulary can supply an off-page verb.

Constrained decoding and deterministic selection. Lexically constrained decoding can enforce target-side requirements during beam search [7]. We apply a stricter product contract: only complete, visible, representable source words are eligible in the early system. The centroid component uses inverse document frequency, embedding similarity, residualization, and maximal marginal relevance (MMR) [8] to add a third word. This component is deterministic but learned-model-dependent because its stronger version uses the trained model's embedding table.

Evaluation under repeated adaptation. Repeatedly adapting to a development set risks overfitting even when individual comparisons are statistically sound [11]. Fine-tuning itself can also be unstable across seeds and hyperparameters [9], and underspecified pipelines may admit many equally plausible but behaviorally different solutions [10]. We therefore emphasize paired comparisons, board separation, sealed tests, and fresh holdouts. The campaign primarily uses an LLM judge, a practice that can scale qualitative evaluation [12] but may introduce position and presentation biases [13]. Identity randomization, exact-string score sharing, auxiliary audits, and explicit claim boundaries reduce but do not eliminate that risk. The overall approach is closer to behavioral product testing [15] than to reference-match evaluation.

3 Task and System Decomposition

3.1 Output contract

The final terminal task maps a bounded task string to a tab title. The Session-11 contract is:

- one to three words and at most 32 characters;
- alphanumeric characters and spaces only;
- Title Case;
- prefer two words, prefer content words present in the task, and name the user's intent rather than the agent interface;
- avoid generic high-frequency titles such as *Code Review*, *Check Status*, or *Fix Bug* unless they are genuinely specific to the task.

Training inputs are synthesized as `title: agent=<a> [cwd=<c>] task=<z>`, using the first paragraph and a bounded character window rather than the entire raw log. Earlier SO-board experiments used Stack Overflow page bodies and emphasized fully visible source words. The difference matters: the early system learned a Stack Overflow title dialect, while the product domain contains commands, project names, conversational fragments, and agent-control language.

3.2 A layered view of the namer

We write the displayed title as

$$\hat{y} = g(x,\ d(p_\theta(\cdot \mid x),\ \mathcal{C}(x)))\,, \tag{1}$$

where x is the task, p_θ is the neural model, $\mathcal{C}(x)$ is a candidate or legality set, d is the decoding or selection rule, and g is deterministic post-processing. This decomposition prevents several common attribution errors. A score change after switching from greedy to beam search is not a weight improvement. A centroid fill is not part of the generator's learned likelihood. A pointer head changes the output structure even when the encoder is frozen. A title-tuned FLAN model is not equivalent to stock FLAN merely because both share the same base architecture.

The principal components were:

12.7M student.
A compact T5 student with a 6,985-token vocabulary, initially distilled from a title-tuned FLAN teacher.

35M B-9500.
A six-encoder/three-decoder T5 trained from scratch on 4.86 million body-to-Stack-Overflow-title pairs, then selectively fine-tuned on terminal titles.

Centroid v2.
A deterministic source-word selector using the trained B-9500 embeddings, inverse document frequency, corpus-mean subtraction, residualization, and MMR.

Hybrid A.
A $2 + 1$ display rule: retain at most two eligible namer words and fill the remaining slot with centroid v2. Older notes sometimes used "Hybrid A" for an entire stack; this paper uses it only for the glue rule.

Dense P1.
A 24.5M inference system that freezes the 22.5M-parameter 6t encoder, discards its autoregressive decoder, and adds a 2.0M-parameter pointer head over the schema `[optional intent] + object + [optional qualifier]`.

Title-SFT FLAN.
An untied 76,961,152-parameter FLAN-T5-small descendant trained to emit free-form task titles directly.

3.3 Two boards, not one leaderboard

The campaign uses two evaluation domains. Their scores must not be averaged or treated as a single staircase. Table 1 records the distinction.

The canonical gaps use different packets and are not cross-board score comparisons. They show that the relative difficulty of the benchmark changed after domain transfer. A rule that helped on SO-board could hurt on Terminal-board; for example, an exact-three-word policy gained 0.39 on one SO-board experiment and lost 0.36 on Terminal-board [21].

3.4 Judging and statistical comparisons

Most packets use Gemini 3.5 Flash-Lite as a blinded pointwise judge. Candidate identities are randomized, identical strings within a row receive the same score, and the rubric assigns contextual usefulness from 1 to 10. Early reports use "useful" at score ≥ 4 and "fail" at score ≤ 2; later product reports use score ≥ 6 as a co-primary "solid title" rate. These thresholds are not merged.

Table 1: Evaluation boards and their claim boundaries. Means are comparable only among systems scored in the same packet.

Board	Input domain	Principal evaluation surfaces	Canonical early gap
SO-board	Stack Overflow question pages	Clean-development 1,000; one consumed sealed 1,000; separate 40- and 200-row Codex audits	FLAN+centroid v2 6.11 vs. 35M pin 5.96, a within-board gap of 0.15
Terminal-board	Redacted or synthesized coding-agent terminal tasks	Terminal-dev 300; reusable-dev 500; sealed-v2 516; confirm-v1 335; two never-seen 1,000-row holdouts; a separate 200-row holdout remains closed	FLAN+centroid v2 7.11 vs. 35M pin 5.94, a within-board gap of 1.17

The number of candidates shown to the judge and the packet composition changed across stages. Absolute means can therefore drift even for a frozen system. We treat within-packet paired differences as primary. Where reported, confidence intervals are paired row bootstrap intervals, typically with 100,000 resamples; win/tie/loss counts provide a distribution-free view of how the mean changed [16].

The four-beam study also used an identity-blind Codex audit with a strict usefulness rubric. This was an auxiliary LLM audit, not a human study. Some later notes include qualitative human spot checks, but the campaign does not contain a powered, systematic human evaluation. The two final 1,000-row holdouts were never seen during the preceding development campaign and are the strongest promotion evidence in this paper.

3.5 Latency and payload

Session 10 measured exported local workers with `onnxruntime-node` 1.24.3 on an Apple M4 Max CPU using one thread, 40 rows, and five discarded warmups. String glue added approximately 0.02 ms. The 35M and pointer clocks are measured ONNX runtimes; the 77M payload is an INT8 estimate and its 29.3 ms clock comes from the Session-10 SFT graph. Importantly, the published 6t quality scores came from fp32 PyTorch decoding while the 10.1 ms clock came from the shipped INT8 ONNX worker. The campaign did not complete a same-packet quality comparison of those artifacts.

4 Campaign Design

The campaign was adaptive. Each result changed the next intervention, and several development packets were deliberately reused. To avoid laundering exploratory iteration into confirmatory evidence, we use four evidence levels:

1. **Exploratory diagnosis**: a result that identifies a possible mechanism or error mode.
2. **Matched packet**: competing systems scored together on the same rows with paired analysis.
3. **Sealed or first-look packet**: configuration frozen before opening a held-out set.
4. **Replicated fresh holdouts**: the same direction on two never-seen samples.

Table 2 condenses the ten public studies. Appendix A provides a fuller ledger and Appendix B records the negative interventions.

Table 2: The staged campaign. "Board" identifies the only leaderboard on which the stated result applies.

Stage	Public note	Board	Intervention	Main conclusion
1	Sniff Test	matched controls	literary embedding initialization and body seeds	A 37% repetition drop in one pair did not recur as a sequence-order mechanism in 31 deterministic runs.
2	Two recipes	SO-board	centroid glue and larger pretrained baseline	Centroid-only beat the 12.7M student; FLAN+centroid led the first product packet.
3	Mid GSG	SO-board	4.86M-pair GSG pretraining and centroid v2	A 35M scratch model tied raw FLAN and closed the glued gap to 0.15.
4	Four beams	SO-board	legal source-word beam-4, frozen weights	Search produced a sealed +0.752 relative win, but absolute Codex usefulness failed the promotion gate.
5	Session 4	Terminal-board	task-aligned FLAN-title SFT for 35M	A small +0.12 gain replicated, but the domain-specific FLAN gap remained large.
6	Techniques log	Terminal-board	six-teacher judged SFT and many follow-ups	Selective labels produced +0.19 on $n = 1000$; self-distill, hinge, free decode, terminal-IDF, and rankers did not add a second bump.
7	176k CE	Terminal-board	175,642 legal Flash-Lite labels	Lower cross-entropy and 86× more labels did not beat the selective judged set.
8	Session 9	Terminal-board	Dense P1 schema pointer	P1 beat 6t by +0.13 on sealed-v2 and was much faster, but remained 0.60 behind title-tuned FLAN.
9	Session 10	Terminal-board	common quality-latency board and FLAN task SFT	Task-tuned FLAN beat P1 by +0.50 and stock FLAN by +4.17 on one packet; fresh confirmation favored raw output over glue.
10	Session 11	Terminal-board	leak-checked continuation on gold plus three teachers	The selected 76.96M model improved the frozen ship by +0.21 and +0.20 on two unseen 1,000-row holdouts.

Table 3: Early SO-board packet ($n = 1000$). All rows were judged together. Useful is score ≥ 4 and fail is score ≤ 2.

System	Mean	Useful (%)	Fail (%)
12.7M student	3.01	31	54
Centroid only	4.21	55	23
Early 35M + centroid	4.72	64	17
Title-tuned FLAN raw	5.22	68	18
Title-tuned FLAN + centroid	5.43	77	10

5 Results

5.1 A conspicuous error reduction was not evidence of the proposed mechanism

The campaign began with an apparent qualitative improvement in the 12.7M student. In a fixed 907-task comparison, replacing the initial embedding state with embeddings exposed to ordered Jules Verne text reduced repeated-token titles from 145 to 92, or from 16.0% to 10.1%. Among titles of at least three tokens, repeats fell from 121 to 70. Yet mean teacher similarity was essentially unchanged: 0.3512 versus 0.3519, a paired difference of -0.00066 whose cluster-bootstrap interval crossed zero [17].

The comparison was confounded because it also changed the random transformer body. A deterministic 31-run matched control crossed three body seeds with three donor seeds and compared ordered Verne with a frequency-matched shuffled version. Ordered Verne produced 14.75% repeated titles versus 13.79% for shuffled text. It was worse in six of nine matched cells, tied once, and better twice; teacher similarity was 1.0% lower. The data therefore did not support a stable literary sequence-order effect.

This study established the campaign's first methodological rule: a visible change in error topology is worth investigating, but one checkpoint pair cannot identify its cause. It also motivated later use of frozen components, paired packets, and independent replications. The repetition improvement itself was real for that pair; the causal story was not.

5.2 Deterministic glue initially carried more value than the tiny generator

On the first SO-board product packet, the 12.7M student scored 3.01 mean, with 31% useful and 54% fail. A deterministic centroid highlighter alone scored 4.21, 55% useful, and 23% fail. The early 35M-plus-centroid system reached 4.72, while raw title-tuned FLAN reached 5.22 and FLAN+centroid reached 5.43 [18]. Table 3 shows the same-packet comparison.

The centroid did not "understand" the task in the generative sense. It selected salient source words under a deterministic heuristic. Nevertheless, it removed a large fraction of catastrophic failures and improved FLAN as well as the small model. In a narrow product, a low-capacity neural generator can be weaker than its deterministic wrapper.

The next 35M checkpoint, B-9500, was trained from scratch on 4.86 million body-to-title pairs. With the old centroid it scored 5.65, tying raw FLAN at 5.64 and beating the earlier 35M stack at 5.10. Centroid v2 then raised FLAN from 5.91 to 6.11 and B-9500 from 5.89 to 5.96 in a six-way packet, leaving a 0.15 gap [19]. Better task-aligned pretraining had made the neural component

competitive, but the deterministic selector remained load-bearing.

5.3 Search changed the ranking of frozen weights, then failed an absolute gate

The decisive SO-board relative gain came from decoding, not training. B-9500 and centroid v2 were frozen. The decoder was changed from greedy output to beam-4 over complete pairs of words that were (i) visibly present inside the true truncated source prefix, (ii) fully mapped by the reduced vocabulary at an actual occurrence, and (iii) representable as a target. This repaired subtle cases where punctuation removal or isolated tokenization created strings not actually available to the encoder.

On two fresh development packets, beam-4 beat corrected greedy by +0.481 and +0.453. On a sealed 1,000-row packet, the resulting 35M stack scored 6.20 versus 5.45 for title-tuned FLAN with the same centroid, a paired difference of +0.752 with a 95% bootstrap interval of [0.611, 0.891] and 603/67/330 wins/ties/losses [20]. All 1,000 outputs passed the lexical and length contract. Every continuation checkpoint evaluated under beam-4, including a conservative 500-step continuation, underperformed unchanged B-9500.

The relative result did not clear the preregistered final gate. An identity-blind Codex audit rated 19 of 40 outputs useful (47.5%), below the required 24 of 40 (60%), although the same audit favored the 35M system over FLAN 18/9/13. A later 200-row development audit rated 80 outputs useful (40.0%, Wilson interval 33.5–46.9%). Thus the study supports a sealed relative ranking on SO-board, not a claim of satisfactory absolute product utility.

This distinction is central. Beam search exposed higher-probability legal titles that greedy decoding missed, but it could not add semantic knowledge absent from the model. A constrained decoder can extract latent competence; it cannot guarantee that the competence is adequate.

5.4 The SO-board win did not transfer to terminal tasks

Moving from Stack Overflow pages to terminal tasks changed the relative gap. On the canonical SO-board packet, FLAN+centroid v2 scored 6.11 and the 35M pin 5.96. On the canonical Terminal-board packet, the same named stacks scored 7.11 and 5.94, respectively [21]. We do not compare the absolute means across boards; we compare the within-board deficits, which grew from 0.15 to 1.17.

The error modes explain the shift. Stack Overflow inputs contain descriptive nouns and conventional titles. Terminal inputs include commands, deictic language, repository-specific names, and short conversational turns. The two-source-word restriction that grounded the 35M model on Stack Overflow often preserved the wrong noun or function word on terminal prompts. The SO-derived IDF table and third-word glue could append words such as *Far*, *Tell*, *Few*, or *Yeah*. The intervention that won one board therefore became a domain-specific production liability.

This domain transfer also changed the interpretation of FLAN. The benchmark was not generic stock FLAN. It was a title-tuned FLAN descendant that had learned the local title dialect. Later Session-10 evaluation made this explicit: stock FLAN scored 3.12, while the task-tuned model scored 7.28 on the same packet.

5.5 Selective judged labels beat much larger weakly filtered supervision

The first terminal-domain fine-tuning pass, r3, used 971 FLAN titles that could be spelled from on-page words and received a judge score of at least 6. It improved the 35M pin from 5.56 to 5.67

on a 300-row confirmation packet, a paired +0.12 with interval [0.04, 0.20]. An earlier independent packet showed +0.07 [0.01, 0.14] [21].

The six-teacher `6t_argmax` recipe extended the same principle. It used 2,040 titles passing a score threshold and preserved the constrained beam and Hybrid A display rule. It improved the pin by +0.12 [0.04, 0.20] on a 300-row packet and by +0.19 [0.14, 0.25] on a later 1,000-row packet. Filtering to tasks of at least five or ten words tied rather than clearly surpassed 6t. Self-distillation, contrastive hinge training, unconstrained decoding, terminal-specific IDF glue, and a local ranker failed to add a second reliable gain [22].

A quantity-focused experiment then labeled 219,800 raw GitHub-Archive-derived rows with Flash-Lite and retained 175,642 legal titles, approximately 86 times the number used by 6t. Four 35M continuations drove cross-entropy from about 6.2 to roughly 1.5; the hottest run reached 0.91. Yet on the same 300-row packet, the best arm scored 6.05 versus 6.16 for 6t and 6.62 for FLAN after applying the same beam-4 and glue. The paired best-arm difference from 6t was −0.112 [23].

This is the clearest evidence in the campaign that training loss and label volume were poor product proxies. The large corpus was legal but less selective, less domain-matched, and generated by one judge-teacher protocol. The smaller set was filtered for judged usefulness. Data quality here includes not only title grammar, but the title dialect, task distribution, teacher diversity, and selection rule.

5.6 A structured pointer improved efficiency but exposed a capacity ceiling

Dense P1 froze the 6t encoder and replaced the autoregressive decoder with an 8 MB pointer head. It scored legal titles of the form `[optional intent] + object + [optional qualifier]`, with 64 intent verbs and on-page object/qualifier candidates. On the first-look sealed-v2 516 packet, P1+ Hybrid A scored 6.43 versus 6.30 for 6t+ Hybrid A, a paired +0.13 [0.04, 0.22]. It remained 0.60 [0.49, 0.70] behind title-tuned FLAN [24].

The structure was efficient. P1 used 24.5M inference parameters (22.5M frozen encoder plus 2.0M head) and later measured 3.88 ms per title. It improved object selection while preventing most hallucinations. However, the legal schema still limited phrasing and intent selection; it often overused *Check* and inherited the encoder's representation ceiling. The same source grounding that made it safe also constrained its title dialect.

Session 10 placed P1, 6t, lightweight pickers, and the FLAN variants in one reusable-dev 500 packet. P1+ Hybrid A scored 6.79; 6t+ Hybrid A scored 6.65; a 378-byte logistic picker reached 6.81; a 1.6 MB neural picker reached 6.83; and the oracle over already-judged P1 and 6t titles reached 7.04. Raw title-SFT FLAN scored 7.28. Its paired advantage was +0.50 [0.36, 0.64] over P1 and +0.64 [0.50, 0.78] over 6t [25].

Figure 1 shows this same-packet quality-latency surface. Stock FLAN is included to demonstrate that capacity alone was not sufficient: it was slower than the local small models and scored only 3.12. Task SFT, not the model name, created the high-quality system.

5.7 Glue was competence-dependent

On the Session-10 ship board, Hybrid A improved title-SFT FLAN from 7.28 to 7.41, a paired +0.13 [0.03, 0.22]. That result alone would suggest retaining the same glue for the larger model. A fresh confirm-v1 set of 335 local tasks reversed the decision: raw SFT scored 6.55, while SFT+ Hybrid A scored 5.78. Qualitative inspection showed that the stronger model already named the

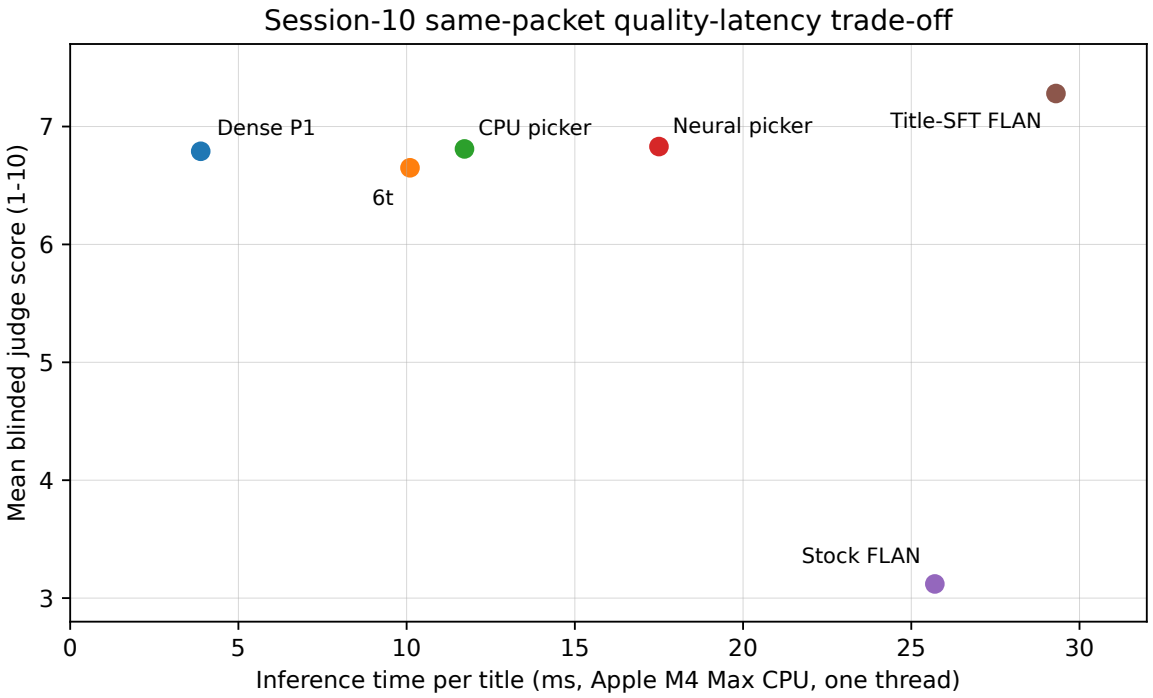

Figure 1: Session-10 reusable-dev 500 quality versus measured local latency. All quality points are from one blinded packet. Latencies are ONNX measurements on an Apple M4 Max CPU with one thread. The published 6t quality and latency refer to different precision artifacts, and the SFT payload is estimated; see Section 7.

job directly and the glue sometimes appended distractors or deleted useful free-form words. The raw model was therefore selected for shipment [25].

The reversal reconciles the earlier findings. Glue is most useful when it repairs a model whose common failure is omission or collapse. It becomes harmful when the base model is already competent and the heuristic's assumptions are weaker than the model's output. A post-processor should be treated as a conditional intervention with its own validation set, not as permanent infrastructure.

5.8 The final continuation replicated on two never-seen holdouts

The final training mix combined production gold titles with three teacher streams: Grok, Claude Opus, and Codex. The historical file contained 7,623 rows, including gold duplicated three times and 6,093 pseudo-labeled rows. Before Session 11, task-hash overlap was checked against terminal-dev 300, holdout 200, both 1,000-row holdouts, reusable-dev 500, confirm-v1, and sealed-v2. The note reports 1,148 task-hash leaks dropped and a retained set of 6,419 rows (372 gold and 6,047 teacher), split into 6,219 training and 200 validation rows [26]. These published counts do not arithmetically reconcile with the stated 7,623-row starting file, so we reproduce them as reported rather than infer an undocumented intermediate filter.

The selected arm, `flat1e4`, continued the untied Session-10 ship for four epochs at learning rate 10^{-4}, batch size 16, seed 1, and flat sample weights. The architecture and tokenizer did not change. Raw greedy decoding used up to 12 new tokens and no Hybrid A glue.

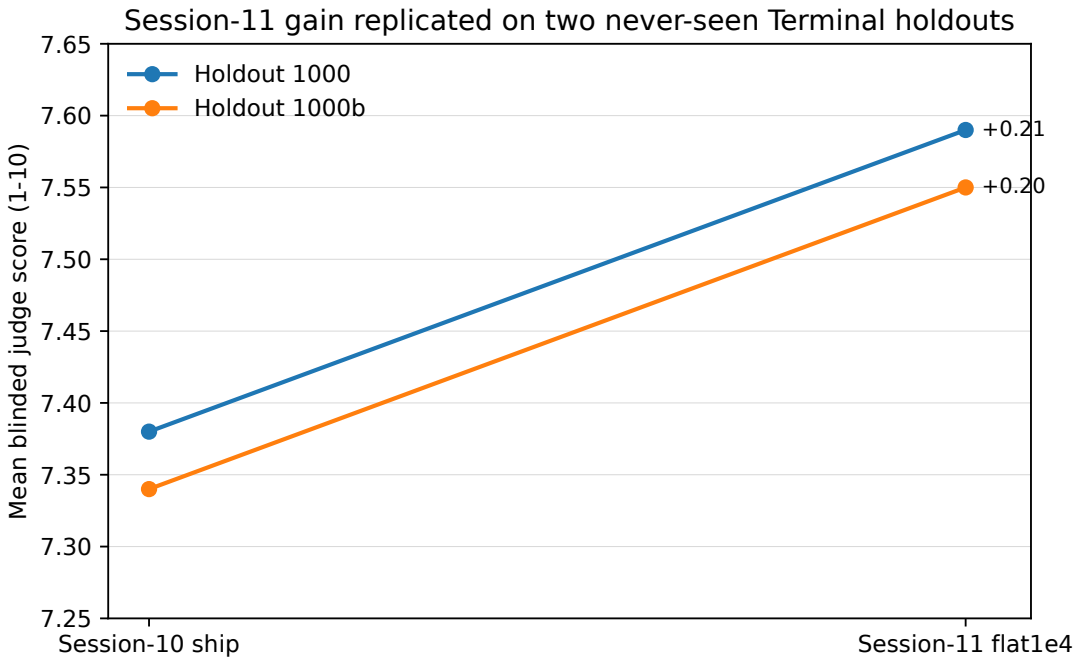

Figure 2: The final continuation improved the frozen prior ship on both never-seen 1,000-row Terminal-board holdouts. These are separate packets and are not averaged with earlier Session-10 development means.

On holdout 1000, `flat1e4` scored 7.59 with 85.7% at or above 6, versus 7.38 and 83.5% for the frozen Session-10 ship. The paired win/tie/loss count was 296/551/153. On holdout 1000b, it scored 7.55 and 84.5% versus 7.34 and 82.5%, with 281/570/149. A reweighted arm was 0.01 higher on the first holdout but fell to 7.51 on the second; `flat1e4` was chosen for two-holdout consistency, more changed titles, and a simpler recipe.

About half of titles were unchanged. The gain came from the changed half, with mean title length increasing from 2.59 to 2.74 words on the second holdout. Examples included *Fix Fix Bug* becoming *Fix SSE Disconnect* and *Apply Deadline* becoming *Apply Target File*; regressions remained, including repeated words. The final result is therefore a modest replicated mean improvement, not a solved task.

The Session-11 graph was not re-benchmarked because the architecture was unchanged. Its inherited product estimate is approximately 193 MB INT8 and 29.3 ms per title, compared with 77 MB and 10.1 ms for the live 6t worker. The project accepted the capacity cost for the quality gain. At publication time the new worker was wired in the development repository, while the public notarized installer still served 6t.

6 Cross-Study Synthesis

6.1 The dominant lever moved as competence increased

The intervention sequence can be read as a competence ladder:

1. The 12.7M model was so weak that a deterministic salience heuristic outperformed it.

2. A task-pretrained 35M model made the neural generator competitive, but decoding and glue still dominated its usefulness.
3. Constrained beam search extracted latent competence from frozen weights and reversed a benchmark ranking on SO-board.
4. Domain transfer exposed missing semantic competence that search could not create.
5. Selective terminal supervision improved the 35M system, and a pointer schema improved its efficiency, but both remained below a larger task-tuned model.
6. Once the 77M model learned the title dialect, the heuristic glue became unnecessary and sometimes harmful; additional leak-checked in-domain supervision produced the final replicated gain.

The same technique therefore should not be assigned a global sign. Beam search was positive for B-9500 but did not rescue continuation checkpoints. Hybrid A was positive for weak generators and negative for the final model on fresh confirmation. Source-word constraints improved grounding but eventually restricted intent expression. The effect of a system component depends on what the model already knows.

6.2 Reference-free product utility was not captured by likelihood

Several objectives diverged:

- teacher similarity stayed flat while repetition changed substantially in the initial 12.7M pair;
- training cross-entropy improved sharply in the 176k experiment while judged title utility fell;
- a sealed relative win over FLAN coexisted with a failed absolute usefulness gate;
- an oracle candidate union reached 7.04, but learned pickers recovered only a small fraction of that headroom;
- stock FLAN had substantial general language capacity yet failed the title dialect until task SFT.

For this task, the target is not a unique reference string. Many titles can be useful, and a high-probability teacher imitation may still be generic or misleading. The product metric must inspect the displayed title in context. Likelihood remains useful for optimization and debugging, but not as the promotion criterion.

6.3 Fresh data mattered more than another sweep

The campaign repeatedly found diminishing returns from training longer on the same material. Beam-era continuation checkpoints lost to the frozen B-9500 model. The 176k corpus did not beat the selective set. In Session 11, first-batch cross-entropy was already about 0.13 because the model had seen the mix before; additional hyperparameter variation produced only small differences. The study's own next-step conclusion was to add new frozen, teacher-labeled rows under the same contract rather than repeat learning-rate and seed sweeps on the same 6.2k rows.

This conclusion is specific to the observed learning curve, but it is consistent with the broader evaluation design. Once a development distribution has guided many decisions, more optimization against it increases adaptive overfitting risk. New training data and new evaluation data both have higher expected information value than another small search over a saturated packet.

6.4 A negative-result registry improved decision quality

The public notes retain superseded and failed interventions rather than presenting only the final staircase. That record prevented repeated work: literary order, self-distillation, free decoding, terminal-IDF glue, generic title CE, and a schema-imitation continuation were explicitly retired. Negative results also sharpened the eventual claim. The final contribution is not "we tried many things and one won"; it is a map of which layer failed under which conditions, with fresh-holdout evidence reserved for the final model.

7 Limitations and Threats to Validity

LLM judge dependence. Gemini 3.5 Flash-Lite is the primary evaluator, and Codex is the auxiliary absolute-usefulness auditor. Both are language models. Identity randomization and paired packets reduce presentation bias, but they do not establish agreement with users. The campaign lacks a powered human study with multiple raters, adjudication, and inter-rater reliability. All absolute utility rates should therefore be read as judge-specific estimates.

Adaptive campaign and reused packets. The work was highly iterative over a short period. Reusable-dev and terminal-dev influenced many decisions. Paired confidence intervals quantify row-sampling uncertainty for a fixed comparison; they do not correct for the full adaptive search over interventions. The two final 1,000-row holdouts reduce this concern for the Session-11 comparison, but not for every earlier mechanistic conclusion.

Private and narrow data. Terminal tasks are drawn from a narrow coding-agent product distribution and are English-centric. Some rows are private, redacted, or employer-identifying and cannot be released. The model may inherit teacher preferences from Grok, Claude Opus, Codex, Gemini Flash-Lite, and earlier FLAN-derived labels. Results may not transfer to other languages, longer titles, general summarization, non-coding tasks, or different agent interfaces.

Board and packet drift. The number of candidates shown to the judge, score thresholds, and packet composition changed across studies. We avoid cross-packet subtraction, but the paper necessarily synthesizes evidence gathered under evolving protocols. The two boards are kept separate; nevertheless, the detailed terminal distributions also evolved from terminal-dev to sealed-v2, confirm-v1, and the final holdouts.

Latency and precision mismatch. Latency was measured on one Apple M4 Max CPU, one thread, and short batches. It may not generalize to other Apple chips, Windows devices, mobile systems, or concurrent workloads. The 6t quality and latency figures refer to fp32 PyTorch and INT8 ONNX artifacts, respectively. The final 77M INT8 payload is estimated, and Session 11 did not re-run latency because its graph was unchanged. Quantized quality was not verified on the same packet.

No population-level seed study. The 31-run matched control is strong enough to reject the original one-pair story as evidence, but it is not a broad population estimate across architectures,

corpus genres, or training recipes. The nine ordered-versus-shuffled matched cells use three body and three donor seeds within one 12.7M architecture.

Model and data release. The public research pages expose recipes, hashes, packet roles, and many aggregate results, but the private task data and some model artifacts are not publicly downloadable. Full external reproduction is therefore not yet possible. A public model card, tokenizer, inference worker, synthetic-data provenance statement, and non-sensitive evaluation sample would materially strengthen a subsequent version.

8 Ethical and Data-Governance Considerations

The terminal corpus can contain sensitive user, repository, or employer context. The campaign used bounded synthesized inputs rather than publishing raw logs, removed secret or employer-identifying rows from qualitative tables, and kept several evaluation sets closed. Session 11 added a mandatory task-hash leakage check against every frozen evaluation surface and removed 1,148 overlapping rows before training.

Closed evaluation protects both privacy and the integrity of future testing, but it reduces independent reproducibility. A responsible release should separate three artifact classes: public model weights and inference code; a non-sensitive synthetic or consented evaluation sample; and private holdouts available only through aggregate evaluation. Teacher-generated labels should retain source-model and filtering metadata because the final model can inherit teacher biases, unsafe naming habits, or memorized strings.

The model is intended only to label local coding sessions. It should not infer sensitive attributes, rank people, or make consequential decisions. A tab title can still reveal project names or task content on screen, so the product should allow users to disable automatic naming, edit titles, and retain local-only inference.

9 Conclusion

A local tab namer is not merely a miniature language model. It is a coupled system of data, weights, candidate constraints, decoding, deterministic glue, runtime artifacts, and evaluation. Across ten public studies, the campaign moved from a confounded seed story to matched controls; from a weak 12.7M student to a competitive 35M stack; from a sealed relative win to an honest failed absolute gate; from Stack Overflow to real terminal-domain failures; from selective judged labels to a failed scale-only synthetic corpus; from autoregressive decoding to a fast pointer head; and finally to a larger task-tuned model whose improvement replicated on two unseen holdouts.

The final selected model is not the smallest or fastest. It is the point at which task-aligned capacity was worth approximately 29.3 ms and an estimated 193 MB INT8 payload. On the strongest available evidence, it improved the prior frozen ship by +0.21 and +0.20 mean judge points on independent 1,000-row terminal holdouts. The broader lesson is methodological: identify which layer changed, compare systems within the same packet, preserve failed gates, separate domains, and reserve fresh data for the claim that matters most.

A Public Study Ledger

Table 4: All ten public study pages synthesized in this paper.

No.	Study	Date	Primary intervention	Claim retained in this paper
1	The Sniff Test [17]	12 Aug	Literary embedding transplant; 31-run matched control	The original repetition drop was pair-specific evidence, not a stable ordered-literature mechanism.
2	Two Recipes, One Glue [18]	14 Aug	Centroid-only and FLAN-plus-centroid baselines	Deterministic salience selection rescued weak systems and improved title-tuned FLAN on SO-board.
3	The 35M that Caught Hybrid A [19]	14 Aug	4.86M-pair scratch GSG; centroid v2	B-9500 tied raw FLAN and came within 0.15 of FLAN+v2 on SO-board.
4	Four Beams, No New Weights [20]	15 Aug	Legal source-word beam-4	Sealed relative +0.752 on SO-board, but the absolute Codex usefulness gate failed.
5	Session 4 [21]	16 Aug	971 selective FLAN titles on terminal tasks	r3 produced a small replicated gain over the 35M pin; FLAN remained far ahead on Terminal-board.
6	Shipping 6t_argmax [22]	16 Aug	Six-teacher judged SFT; negative intervention registry	6t improved the pin by +0.19 on $n = 1000$; many obvious follow-ups did not help.
7	176k Flash-Lite CE [23]	16 Aug	Large single-protocol pseudo-label corpus	175,642 legal labels and much lower CE did not improve judged terminal utility.
8	Session 9 [24]	17 Aug	Dense P1 schema pointer	The pointer head improved quality and speed over 6t but stayed behind FLAN.
9	Session 10 [25]	17 Aug	Common ship board; title-SFT FLAN; timing	Task SFT created the best quality system; raw output generalized better than heuristic glue.
10	Session 11 [26]	17 Aug	Leak-checked continued SFT on multi-teacher mix	The final arm improved the frozen ship on two never-seen 1,000-row holdouts.

B Negative-Result Registry

Table 5: Interventions that were explicitly retired or bounded by the campaign. Deltas are reported only where the source note published a same-packet comparison.

Intervention	Outcome	Interpretation
Ordered literary pre-initialization	Ordered worse in 6/9 matched cells	One lucky pair did not identify a sequence-order mechanism.
Beam-era continuation training	All scored continuations below frozen B-9500	Better likelihood after a decoder change did not imply better beam outputs.
Self-distill best-of-K	Approximately zero	The model's own proposal set did not add a better title dialect.
Contrastive hinge	Only +0.04 vs. 6t	Increasing a likelihood gap did not reliably change beam selection.
Unconstrained/off-page decode	−0.89 in the techniques log	Source-word grounding was protecting the 35M model from hallucination.
Terminal-IDF centroid	−0.19 vs. 6t	Replacing the SO-derived table did not fix the glue; the third word remained load-bearing.
Local ranker/proposer	−0.13 in tested form	Candidate oracles existed, but learned pickers could not recover the teacher-only titles.
175,642-label CE continuation	Best arm −0.112 vs. 6t	Quantity and cross-entropy were weaker signals than selective judged supervision.
Agent1m generic title CE	−0.15 vs. 6t on Session 10	More weakly matched title CE did not improve the local system.
Keep-namer / forced glue variants	Inconsistent; severe fresh-set drops	A fixed heuristic should not be inherited across model competence levels.
Session-9 schema imitation for 77M	Lost on both final holdouts	Teaching the larger model the smaller model's two-word dialect was regressive.

C Reproducibility and Claim-Boundary Checklist

1. **Boards:** SO-board and Terminal-board remain separate. No cross-board means are averaged.
2. **Packets:** Means are compared only within a shared judge packet. Context rows copied from another packet are labeled as such.
3. **Judges:** Flash-Lite is the primary LLM judge; Codex is an auxiliary LLM auditor. Neither is described as human evaluation.
4. **Thresholds:** Early useful ≥ 4 and later solid ≥ 6 rates are not merged.
5. **Final evidence:** The main final claim rests on two never-seen 1,000-row holdouts. The separate terminal holdout 200 remains closed.
6. **Leakage:** Session 11 removed task-hash overlap against all named frozen evaluation sets before training.
7. **Artifacts:** The final model must be loaded with untied output weights (`tie_word_embeddings=false`); tying them yields invalid decoding.
8. **Latency:** Session-11 latency is inherited from the unchanged Session-10 graph, not re-measured. The 6t quality/latency precision mismatch remains unresolved.
9. **Privacy:** Raw private terminal prompts and employer-identifying examples are not released.

10. **Scope:** The conclusions apply to short English coding-session titles under the described product contract, not to general summarization or open-ended generation.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[2] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[3] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[4] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339, 2020.

[5] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[6] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[7] M. Post and D. Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of NAACL-HLT*, pages 1314–1324, 2018.

[8] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, pages 335–336, 1998.

[9] M. Mosbach, M. Andriushchenko, and D. Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2021.

[10] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

[11] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[12] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

[13] P. Wang, L. Li, L. Chen, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[14] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13484–13508, 2023.

[15] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 2020.

[16] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, 2004.

[17] M. I. Hossain. The Sniff Test: A seed lottery in a 12.7M-parameter language model. PorkiCoder Research, 12 August 2026. https://porkicoder.com/research/the-sniff-test.html.

[18] M. I. Hossain. Two tab-title recipes: 16 GB Mac and phone-class silicon. PorkiCoder Research, 14 August 2026. https://porkicoder.com/research/tab-titles-flan-centroid.html.

[19] M. I. Hossain. Mid-pack GSG: The 35M that beat Hybrid A and tied raw FLAN. PorkiCoder Research, 14 August 2026. https://porkicoder.com/research/tab-namer-mid-gsg.html.

[20] M. I. Hossain. Four beams, no new weights: A sealed test confirmed the FLAN win but missed the final gate. PorkiCoder Research, 15 August 2026. https://porkicoder.com/research/tab-namer-four-beams.html.

[21] M. I. Hossain. Session 4: Beating our own namer, still chasing FLAN. PorkiCoder Research, 16 August 2026. https://porkicoder.com/research/session4_results.html.

[22] M. I. Hossain. Shipping 6t_argmax: A techniques log for Terminal-board. PorkiCoder Research, 16 August 2026. https://porkicoder.com/research/tab-namer-techniques.html.

[23] M. I. Hossain. 176k Flash-Lite CE test: Terminal-board, still shipping 6t_argmax. PorkiCoder Research, 16 August 2026. https://porkicoder.com/research/2026-08-16_g300k-flashlite-test.html.

[24] M. I. Hossain. Session 9: Dense P1 + Hybrid A beats 6t, still chasing FLAN. PorkiCoder Research, 17 August 2026. https://porkicoder.com/research/session9_results.html.

[25] M. I. Hossain. Session 10: Every namer, quality versus milliseconds. PorkiCoder Research, 17 August 2026. https://porkicoder.com/research/session10_results.html.

[26] M. I. Hossain. Session 11: More epochs on the Aug-12 mix is the new ship. PorkiCoder Research, 17 August 2026. https://porkicoder.com/research/session11_results.html.