Behavioral Distillation: Internalizing Global Structure through Interaction Consistency

Conrad Kramer

1 Introduction

Behavioral distillation focuses on transferring the interactive dynamics of a teacher model to a student, rather than just final output probabilities. Here we analyze the stability and autonomy of models trained via Wave-Density Attention (WDA) during and after teacher detachment.

1.1 Quantitative Detachment Analysis

Figure **??** presents a quantitative analysis of detachment dynamics. The plot shows student cross-entropy loss versus training steps, highlighting distinct phases where the teacher is active (guidance on) and inactive (guidance off). During the teacher-on phase, behavioral loss decreases steadily, aligning the student to the teacher's attention patterns. Upon detachment, the teacher is removed, yet the student maintains and further improves cross-entropy performance independently, demonstrating successful internalization of behavioral intelligence.

In a representative SmolLM-135M run, the teacher-free student (WDA-only, $\alpha = 1$, teacher taken to zero) improves its evaluation CE over time despite the teacher being absent. Using the same evaluation script and data mix, student CE decreases from ≈ 5.93 (PPL ≈ 376) at an earlier checkpoint to ≈ 4.48 (PPL ≈ 88) and later to ≈ 3.24 (PPL ≈ 25.5). A teacher-only baseline on the same evaluation reports CE ≈ 2.42 (PPL ≈ 11.3). These results show continued post-detachment improvement rather than collapse.

Notably, the post-detachment curve was still trending downward at the final checkpoint; training was halted due to hardware and time constraints (single RTX 3090, weeks-long run), not because of convergence. The sustained decline suggests additional gains were likely with more compute, reinforcing that the detached student was still actively refining behavior rather than stabilizing early.

2 Failure Modes

Behavioral distillation can fail under certain conditions. Insufficient bandwidth in behavioral signals may prevent the student from capturing critical interaction dynamics. Premature detachment before behavioral alignment stabilizes often leads to collapse or degraded performance. Additionally, architectural inductive biases incompatible with the teacher's dynamics can hinder effective transfer, necessitating careful design and tuning.

We also observe practical failure modes: (i) learning-rate plateaus where low LR stabilizes but fails to escape flat regions, requiring brief high-LR "bursts" to break through; (ii) limited data variety, which can stall CE even under teacher-free training; and (iii) prompt-format mismatch in instruction tuning (e.g., training on "User: ... Assistant: ..." while prompting without the assistant cue), which yields on-topic but misaligned completions.

3 Limitations

Our current experiments focus on small- to mid-scale models and a single attention replacement. Kernel choice (triangle vs. sine waves), mask/wave capacity, and gate temperature materially affect both speed and stability, and may require additional fine-tuning when changed. Due to single-GPU constraints, sequence lengths were capped (typically 1024–2048 tokens) and batch sizes kept small (1–4), introducing noisier gradients and slower convergence. Training was also halted before full convergence in the teacher-free phase; the observed continued CE improvement suggests that longer runs, larger batches, and extended contexts would likely yield further gains. Scaling behavior, downstream task performance, and formal guarantees remain open questions.

4 Implications

Behavioral distillation enables:

- Rapid experimentation with new attention mechanisms
- Reuse of pretrained weights across architectures
- Significant reductions in training compute

This opens the door to modular, plug-and-play model design where intelligence is preserved even as structure evolves.

5 Experiments

5.1 Setup

We evaluate on a SmolLM-135M teacher and a WDA-based student of comparable size. Training proceeds in stages: a teacher-on phase to align behavior, followed by a teacher-free detachment phase where the student trains on task loss alone. Phase-1 uses residual-energy behavioral targets to stabilize transfer; Phase-2 removes the teacher entirely; Phase-3 performs instruction SFT on a 100k UltraChat subset. The pretraining data mix includes local JSONL shards from Cosmopedia-v2, FineWeb-Edu-Dedup, Wikipedia, OpenWebText, and Python-Edu, with optional synthetic sine-wave routing to stabilize early dynamics.

5.2 Metrics

We report attention-mask cross-entropy (CE) on a held-out evaluation split and perplexity (PPL), with

$$\mathrm{PPL} = \exp(\mathrm{CE}).$$

We also track routing diagnostics (active masks and entropy) to ensure WDA is engaged rather than collapsed.

5.3 Main Results

Table 1 summarizes teacher-free progress after detachment. The student continues to improve without the teacher, indicating that behavioral alignment is retained and exploited under task loss alone.

Model / checkpoint	CE ↓	PPL ↓
Teacher-only baseline	2.42	11.3
Student (earlier, teacher-free)	5.93	376
Student (mid, teacher-free)	4.48	88.4
Student (late, teacher-free)†	3.24	25.5

Table 1: Teacher-free student continues to improve after detachment on the same evaluation mix. † Training halted due to compute limits while loss was still decreasing.

5.4 Ablations and Diagnostics

Qualitative ablations reveal consistent patterns:

- **Behavioral target**: residual-energy distillation stabilizes early transfer and avoids loss explosions compared to feature-level matching.
- **LR bursts**: brief high-LR stages break CE plateaus that otherwise persist under teacher-free training.
- **Data variety**: adding Wikipedia/OpenWebText shards improves CE trends relative to narrower mixtures.
- **Routing health**: typical runs sustain active masks ≈ 24 with entropy ≈ 3.17, indicating non-degenerate WDA usage.
- **Instruction format**: SFT requires prompt-format alignment ("User: ... Assistant: ...") for consistent, on-topic completions.

6 Why This Changes Scaling Laws

Traditional scaling laws posit that achieving state-of-the-art performance requires massive compute investment, often involving retraining on trillions of tokens. Behavioral distillation alters this paradigm by decoupling intelligence acquisition from architecture instantiation. Instead of retraining from scratch, behavioral distillation leverages pretrained teacher models as reservoirs of learned dynamics and can substantially reduce the additional compute needed to instantiate a new architecture. This enables faster iteration and architectural innovation without prohibitive resource expenditure.

Reviewer Rebuttal (Preemptive)

Is this just feature matching? No. Feature matching enforces representational similarity at fixed layers. Behavioral distillation enforces dynamic similarity: how attention allocates context, how signals propagate through depth, and how residual energy evolves. The student is not constrained to reproduce teacher activations, only their functional usage.

Is the teacher still required at inference? No. The teacher is fully removed after detachment. Post-detachment training and evaluation are performed with the student alone.

Does this trivially reduce to standard distillation? No logits, KL terms, or probability matching are used. The transfer target is internal behavior, not output distributions.

Why is WDA necessary? WDA exposes a continuous, low-dimensional control surface over attention behavior, making it particularly amenable to behavioral alignment and subsequent autonomous operation.

7 Appendix: Behavior vs Representation

A common objection to non-logit distillation methods is that they reduce to feature or representation matching under a different name. This appendix clarifies why behavioral distillation is fundamentally distinct.

Representation matching constrains a student to approximate the teacher's internal activations at fixed layers. Such objectives implicitly assume that intelligence is localized in static feature vectors. In contrast, behavioral distillation targets dynamic properties: how attention mass is allocated over context, how signals are routed across layers, and how residual energy evolves during inference.

Two models may share no representational similarity while exhibiting equivalent behavior. An analogy is control systems: distinct internal states can implement the same input–output policy if their dynamics are aligned. Behavioral distillation enforces alignment at the level of policy execution rather than state realization.

Wave Density Attention is particularly well-suited to this regime because it exposes a low-dimensional, continuous control surface over attention behavior. Supervising this surface aligns how the model uses its representations, not what those representations are.

Empirically, this distinction is validated by teacher detachment. After behavioral loss is annealed to zero and the teacher is removed, the student continues to learn under task loss alone. Feature-matching methods typically collapse under such removal, as the supervision target disappears. Behavioral distillation persists because the learned behavior has been internalized into the student's own dynamics.

In summary, representation similarity is neither necessary nor sufficient for intelligence transfer. Behavioral alignment is sufficient, and Wave Density Attention provides a concrete mechanism by which such alignment can be learned and retained.

8 Related Work

Prior work on knowledge distillation relies on logit matching or intermediate feature regression. Our approach differs by targeting attention behavior as the primary transferable substrate, aligning with emerging views that model intelligence is encoded in interaction dynamics rather than static parameters.

9 Conclusion

We have demonstrated that Wave Density Attention supports behavioral distillation, allowing intelligence transfer without logit dependence. Continued post-detachment improvement, halted only by compute limits, highlights the potential for further gains with modest additional resources. This represents a step toward more flexible, efficient, and interpretable model training paradigms.

Reproducibility

All experiments were conducted using a single training pipeline with feature-gated ablations. Detachment curves are produced via `wave_dencity/scripts/train_transplant.py` and evaluated with `wave_dencity/evals/eval_transplant.py` (attention-mask CE). Instruction SFT uses a 100k UltraChat subset formatted as "User: ... Assistant: ..." (with a newline between turns). Due to single-GPU limits (RTX 3090, 24 GB VRAM), training used reduced sequence lengths and small batch sizes with gradient accumulation; longer multi-GPU runs would allow full-context training and cleaner convergence. Code and configurations will be released upon publication.