Product of Experts as Scalable Local Learning: Modular Construction at 1.3B Parameters

Jaepil Jeong
Cognica, Inc.
Email: jaepil@cognica.io
Date: April 19, 2026

"The idea of combining the opinions of multiple different 'experts' is very old, but it has generally been implemented by averaging the probability distributions produced by the individual experts. This corresponds to using a mixture model. It is also possible to combine experts multiplicatively..."
— Geoffrey E. Hinton, *Training Products of Experts by Minimizing Contrastive Divergence*, 2002

Abstract

Backpropagation requires global state: a single loss propagated through all layers, activations retained across the network, gradient synchronization across parameters. We show that the Product of Experts (PoE) framework provides a principled foundation for local learning eliminating this dependency, and validates at 1.3B production scale that modular model construction — training-time modularity, inference-time composability, post-hoc extension, module preservation under specialist SFT, quality-positive composition — is realizable.

Quality trajectory. Clustered PoE on 1.3B GPT against matched BP shows 6.0% BPB gap at r=10 and **6.52% gap at r=20 from-scratch** (final: BP 0.676788, PoE 0.720935) — weak compression (~12% relative, consistent with the r=10 → r=20 projection). Gap *widens convexly* within the r=20 run: 4.3% at step 1k → 5.0% at step 15k → 5.8% at step 20k → **6.52% at step 26,430 (final)**, with 31% of the total widening concentrated in the final 6K warmdown steps. The widening pattern — and its acceleration through warmdown — shows BP's global gradient coordination delivers its largest advantage during fine-grained optimization, where PoE's per-stage local gradients cannot match global adjustment precision. Combined with the weak r=10 → r=20 compression, the evidence supports a **structural-floor interpretation** (H-S): the gap reflects a bounded architectural cost of local learning, not a training-budget artifact that vanishes with scale. An r=30 from-scratch experiment remains the direct test of whether an additional slow-compressing H-E component exists. On 22-task CORE, gap is task-polarized rather than uniform: PoE underperforms on rare-fact (Jeopardy −81%, SQuAD −44%) but exceeds BP on commonsense (PIQA +5.0pp, CSQA +5.8pp) and algorithmic patterns (BigBench CS Algorithms +11.4pp).

Architectural properties. Stage 1 alone (25% compute) achieves 87.5% of full-model factual accuracy; WAND adaptive pruning delivers 1.82× wall-clock at 100% top-1 agreement; Stage 1 as natural drafter for speculative decoding yields 1.87× at 88% acceptance; post-hoc specialists attach to frozen base without retraining. All transfer to Apple Silicon via MLX with framework-agnostic consistency.

Modularity validated on six pillars. (1) Training-time modularity via detached per-stage losses. (2) Inference-time composability — same checkpoint serves multiple compute tiers. (3) Post-hoc specialists without retraining. (4) **Base preservation under SFT**: compute-matched 5557-step dual-head SFT leaves Stages 1–4 bit-identical (delta = 0.0000 across twelve checkpoints, Washington rank 1 / logit +9.81), while v2 single-head destroys the same path by −14.73 logit. (5) **Composable quality gains**: log-space parallel composition $\{[1..4], [1..5]\}$ strengthens factual-retrieval margins by ~2.4 logit over strongest single-branch — PoE algebra producing sharper joints. (6) **Heterogeneous specialist ecosystem**: 2-layer SQuAD factual on same base produces orthogonal retrieval-trust (4–6× less gullible on wrong-RAG than 6-layer chat specialist), demonstrating task-dependent depth and SFT-data causality.

Deployment position. On-device inference is PoE's natural habitat: memory, battery, device heterogeneity align with prefix pruning, WAND, speculative decoding, elastic depth, modular update. The rare-fact weakness is the structurally-predicted consequence of locality — the same constraint producing the same profile in biological cognition, which has converged on external retrieval. §8.2.1 reports multi-token greedy with a retrieval passage reaching $P(\text{George Washington}) = 0.43$ on an otherwise-unreachable ambiguous query (712×). The quality gap is measurable,

bounded, and stable — a clean architectural trade-off rather than a failure mode.

---

1. Introduction

1.1 The Global State Problem

BP (Rumelhart et al., 1986) maintains global state: **(G1)** single scalar loss at the output; **(G2)** all activations $h_0, \ldots, h_{L-1}$ stored until backward; **(G3)** gradient for layer k depends on layers $k+1, \ldots, L$; **(G4)** in distributed training, all-reduce across all workers before any update.

Each global dependency creates a scaling bottleneck:

Global State	Bottleneck	Consequence
(G1) Single loss	Sequential backward	Pipeline bubbles (~50% waste)
(G2) All activations	Memory	$O(L \cdot B \cdot T \cdot d)$ peak
(G3) Cross-layer gradients	Cascading instability	Loss spikes, gradient explosion
(G4) All-reduce	Communication	Scaling wall at thousands of GPUs

1.2 Local Learning as the Solution

Local learning provides each layer an independent learning signal, removing global coordination. Prior: target propagation (Lee et al. 2015), greedy layerwise (Belilovsky et al. 2019), Forward-Forward (Hinton 2022), HSIC (Ma et al. 2020), decoupled greedy (Löwe et al. 2019), synthetic gradients (Jaderberg et al. 2017). These demonstrate viability on small benchmarks but incur 5–15% accuracy penalties; few have been validated across architectures or at scale.

1.3 The Present Contribution

We establish that the Product of Experts (PoE) framework (Hinton, 2002) provides a unified foundation for local learning from small networks to production-scale LLMs. The central thesis: **modular model construction is empirically realizable at 1.3B production scale**, validated along six independent axes:

1. **Cross-architecture validation** (Appendix A): PoE works across MLPs, CNNs, ResNets, Transformers with gaps of 0–12%.
2. **Systems-level properties** (§4): Layer independence enables lossless prefix prediction, reduced-bubble pipelines, elastic scaling, parallel branching — provably unattainable under BP.
3. **Production-scale empirical validation** (§5): Clustered PoE on 1.3B GPT shows a **bounded BPB gap** (6.0% at r=10, **6.52% at r=20 from-scratch** — ~12% weak compression). The r=20 run's *within-run* trajectory shows gap widening convexly through warmdown (4.3% at 1k → 5.8% at 20k → 6.52% at 26,430 final), with 31% of the total widening concentrated in the final 6K warmdown steps. BP's global coordination provides its largest advantage in fine-grained optimization where PoE's per-stage local gradients cannot match. Task-polarized CORE profile. WAND and speculative decoding deliver ~1.85× speedups. Evidence supports **structural-floor interpretation (H-S)**; r=30 from-scratch training (§10.1) remains the direct test of any slow-compressing H-E component.
4. **Post-hoc modular extension and preservation** (§6): Specialist stages attach to the frozen base without retraining. Compute-matched dual-head SFT (§6.5) preserves Stage 4's factual path bit-identically (delta = 0.0000 across twelve checkpoints, Washington rank 1 / logit +9.81), while v2 destroys it by −14.73 logit units.
5. **Composable quality gains** (§6.5): Log-space parallel composition $\{[1..4], [1..5]\}$ strengthens factual-retrieval margins by +2.4 logit units over the strongest single-branch — a direct instance of PoE algebra. Branch weights give an inference-time tuning axis between absolute-confidence and margin-robustness without retraining.
6. **Heterogeneous specialist ecosystem** (§6.6): A 2-layer SQuAD-trained specialist on the same base produces an

orthogonal retrieval-trust profile (4–6× less gullible on wrong-RAG than the 6-layer chat specialist). Causal test of SFT-data-induced inductive bias; fourth orthogonal axis of base-head preservation; specialist depth is task-dependent.

7. **On-device verification** (§7): Apple Silicon MLX benchmarks confirm speedups transfer to consumer hardware with framework-agnostic consistency.

Modularity is not a concession forced by scale limits but a quality-positive architectural choice: independently-trained modules can be preserved, composed, and weighted at inference to produce retrieval quality exceeding the base model. We further propose a modular-update discipline (§10.4) in which base-model version changes behave like software framework version transitions.

1.4 Notation

Symbol	Definition
z_k	Logit output of expert k
$f_k(\mathbf{y} \mid \mathbf{x})$	Expert k's predictive distribution
Z	Partition function (normalizing constant)
K	Number of experts (layers or stages with classification heads)
S	Stage size (layers per stage in clustered PoE)
$H(p)$	Shannon entropy: $-\sum p_i \log p_i$
CE	Cross-entropy loss
PPL	Perplexity: $\exp(\text{CE loss})$
P	Number of pipeline stages
M	Number of micro-batches per training step
BPB	Bits per byte (tokenization-invariant evaluation metric)

2. Background and Mathematical Foundation

2.1 Product of Experts

Definition 2.1 (Product of Experts; Hinton, 2002). Given K expert distributions $f_1, \ldots, f_K$ over the same output space, the PoE combines them multiplicatively:

$$p_{\text{PoE}}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{k=1}^{K} f_k(\mathbf{y} \mid \mathbf{x}) \tag{1}$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_k f_k(\mathbf{y} \mid \mathbf{x})$ is the partition function.

Proposition 2.2 (Log-Space Equivalence). For softmax experts with logits $z_k \in \mathbb{R}^C$, the PoE combination reduces to logit summation:

$$p_{\text{PoE}}(\mathbf{y} \mid \mathbf{x}) = \text{softmax}\left(\sum_{k=1}^{K} z_k\right) \tag{2}$$

Proof. Each expert $f_k(\mathbf{y} \mid \mathbf{x}) = \text{softmax}(z_k) \propto \exp(z_k)$. The product $\prod_k \exp(z_k) = \exp(\sum_k z_k)$. Normalization yields $\text{softmax}(\sum_k z_k)$. □

Remark 2.2.1 (PoE as Log-Opinion Pool). The log-space combination in Proposition 2.2 is the normalized Logarithmic Opinion Pool (Log-OP) from Bayesian evidence combination. §2.4 anchors four theorems from Jeong (2026b) that place clustered PoE in a broader framework; production instantiations in §5.3 (WAND), §6.5.5 (parallel composition), §8.4 (multi-head at output layer).

2.2 Flat vs Hierarchical Factorization

Definition 2.3 (Flat PoE). Detached activations sever the gradient path between layers:

$$h_k = \text{Block}_k(\text{sg}(h_{k-1})), \quad \text{logits}_k = W_{\text{head}} \cdot \text{norm}(h_k), \quad \mathcal{L}_k = \text{CE}(\text{logits}_k, y) \tag{3}$$

Definition 2.4 (Hierarchical PoE). Restores the conditional forward path (no detach between layers) while maintaining independent per-head losses. Layer l receives gradients from $\mathcal{L}_l, \mathcal{L}_{l+1}, \ldots, \mathcal{L}_L$ via the forward path.

Definition 2.5 (Clustered PoE). Groups layers into stages of S layers: intra-stage full BP; inter-stage gradient detachment; per-stage CE via shared head. For L layers producing $K = \lceil L/S \rceil$ stages:

$$\mathcal{L}_{\text{clustered}} = \frac{1}{K} \sum_{k=1}^{K} \text{CE}(W_{\text{head}} \cdot \text{norm}(h_{kS}), \mathbf{y}) \tag{4}$$

2.3 Entropy-Based Shrinkage Gating

Definition 2.6 (Shrinkage Weight). For expert k with logits z_k:

$$w_k(\mathbf{x}) = 1 - \frac{H(\text{softmax}(z_k))}{\log C} \tag{5}$$

High-confidence experts (low entropy) receive weight ≈ 1; uncertain experts ≈ 0.

Principle 2.7 (Training-Inference Separation; Empirical). Training-time shrinkage contaminates likelihood with an inference-time prior; optimal estimation follows $\text{posterior} = \text{likelihood} \times \text{prior}$ with terms estimated independently. Retroactive (inference-only) shrinkage outperforms training-time by +4.3pp on Split-CIFAR-10 (Appendix A.3). Empirical principle, not theorem.

2.4 Bayesian Foundation: Companion Paper Framework

The §2.1–§2.3 PoE framework inherits an analytical derivation from Jeong (2026b), which derives NN structure from first-principles Bayesian inference over multiple relevance signals. Four results anchor the present paper:

Theorem 2.4.1 (Attention as Log-Opinion Pool; Jeong 2026b, Theorem 8.3). Given n calibrated probability models P_i, each representing an independent estimate of relevance, the attention-weighted aggregation in log-odds space,

$$S = \sum_{i=1}^{n} w_i(q, s_i) \cdot \text{logit}(P_i), \quad w_i \geq 0, \ \sum_i w_i = 1, \tag{6}$$

is mathematically equivalent to a normalized Logarithmic Opinion Pool:

$$P_{\text{Log-OP}} = \sigma\left(\sum_{i=1}^{n} w_i \, \text{logit}(P_i) \right). \tag{7}$$

This is in turn mathematically equivalent to Hinton's Product of Experts (Hinton, 2002): experts' distributions are multiplied with normalization, and in the logit domain the product becomes an exact linear combination. The attention weights $w_i(q, s_i)$ are therefore the context-dependent exponents in a PoE ensemble.

Remark 2.4.2 (Multi-head as parallel PoE aggregators; Jeong 2026b, Remark 8.6). The progression from a single-aggregator PoE to multi-head attention is a probabilistic generalization: each head independently performs Log-OP aggregation with its own context-dependent reliability weights, and the heads' outputs are combined additively. Multi-head attention is an ensemble of parallel PoE aggregators.

Theorem 2.4.3 (WAND as exact neural pruning; Jeong 2026b, Theorem 8.7.1). Consider a Log-OP aggregation $S = \sum_i w_i v_i$ in which each value v_i admits a computable upper bound $\mathrm{ub}(v_i) \geq v_i$. A token (or stage, or head) i can be **exactly pruned** — skipped without affecting the top-k output — when

$$\sum_{j \in \mathcal{A}} w_j v_j + \sum_{j \notin \mathcal{A}} w_j \cdot \mathrm{ub}(v_j) < \theta, \tag{8}$$

where $\mathcal{A}$ is the set of already-evaluated components and θ is the current k-th highest aggregated score. This is the WAND pruning condition from information retrieval, transferred to the Log-OP / PoE setting. Corollary 8.7.2 of the companion paper extends the result to Block-Max WAND for head-level pruning in multi-head architectures.

Theorem 2.4.4 (Log-OP sharpening; Jeong 2026b, Remark 8.3.1). Log-OP averages *logits* and produces a *product distribution*: the combined model sharpens by requiring consensus among experts. Probability mass concentrates on hypotheses supported by multiple sources, and dissipates where only one source supports. This structural property — not shared by Bayesian Model Averaging, which averages probability values and produces a mixture distribution — is why attention is empirically effective at focusing on contextually relevant information, and why independent experts with complementary evidence produce a sharper joint distribution than any single expert alone.

How these anchor the paper. §§5–6 findings are production-scale instantiations: **per-stage detachment (§3.1)** realizes Log-OP at stage level (Theorem 2.4.1); **WAND stage pruning (§5.3)** realizes Theorem 2.4.3 — p99-bounded per-stage deltas yield exact top-1 preservation at 1.82×; **dual-head (§6.5)** extends Remark 2.4.2 to output projection — base = frozen aggregator over Stages 1–4, specialist = additional expert; **parallel composition (§6.5.5)** is a direct Theorem 2.4.4 instance — two independently-trained branches combine via log-space product for +2.4 logit margin. Jeong 2026d further establishes effective rank as the dominant mechanism governing representational capacity; the Transformer residual stream provides bandwidth-dampening absent in CNNs — supports the nanochat-inherited x_0 blending and ResFormer value embeddings.

Broader pattern: structures as implicit statistical models. Inverted index is an implicit statistical model whose statistics support BM25 as a Bayesian likelihood ratio (Jeong 2026a); ANN indexes are statistical models of embedding-space density (Jeong 2026c). Clustered PoE makes this explicit at the stage level — each stage trained as an independent predictor through its own CE produces a calibrated signal for Log-OP combination. Stage 1's 87.5% full-model factual accuracy is the empirical signature.

The rest of this paper treats Theorems 2.4.1–2.4.4 as established background and focuses on empirical validation at 1.3B.

2.5 Related Work

Local learning. Greedy layer-wise pretraining (Bengio et al. 2007; Hinton et al. 2006) fell out of use as end-to-end BP proved more effective. Recent: target propagation (Lee 2015), Forward-Forward (Hinton 2022), decoupled greedy (Belilovsky 2019; Löwe 2019), synthetic gradients (Jaderberg 2017), HSIC (Ma 2020) — 5–15% accuracy gaps on small-to-medium benchmarks; few validated across architectures or scaled to 1B+.

MoE and adapters. Token-level MoE (Shazeer 2017; Fedus 2022; DeepSeek-V2/V3) routes tokens through FFN expert subsets. Adapters (Houlsby 2019; LoRA, Hu 2022) add deltas to frozen backbones. Universal Transformers (Dehghani 2019) share parameters recurrently — orthogonal to clustered PoE's independent per-stage parameters with disjoint losses. §6 discusses stage-level MoE as a qualitatively distinct primitive.

On-device LLM inference. Quantization (GPTQ, AWQ), distillation (DistilBERT), runtimes (llama.cpp, MLX, Core ML, ExecuTorch) treat on-device as a compression target. This work proposes a different path: architecture designed at training time with on-device properties as intrinsic features.

3. Method: Clustered Local Learning for LLMs

3.1 From Per-Layer to Per-Stage

Per-layer PoE at GPT-2 scale incurs 12% quality gap and 4.4× training slowdown (vocabulary projection dominates compute in small models; Appendix A.2). Two design pressures motivate clustered PoE: **(i) compute efficiency** — with K stages instead of L layers, only K vocabulary projections are needed, reducing overhead to 1.33× at $S = 5$; **(ii) quality headroom from intra-stage coordination** — full BP within a stage preserves fine-grained gradient flow over short dependency ranges, while inter-stage detachment preserves systems benefits (§4).

Figure 3.1 (Standard BP training). Single CE loss at final `lm_head` propagates gradients through every stage in reverse — unbroken chain from CE loss back to Stage 1.

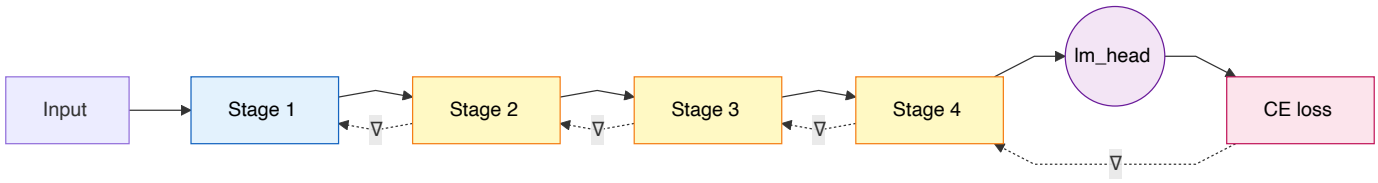
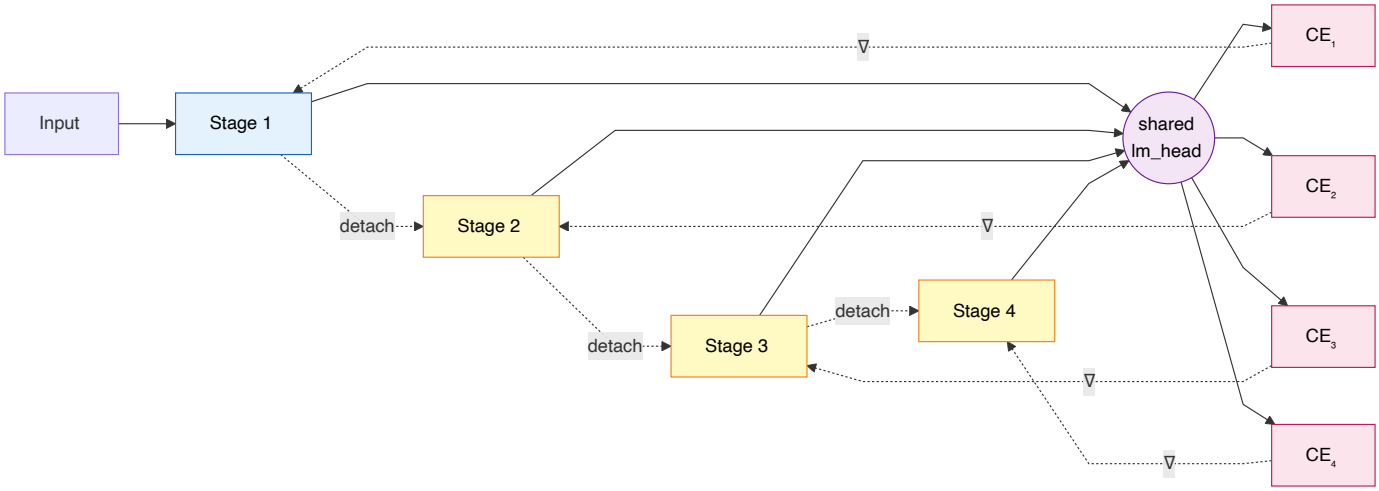


Figure 3.2 (Clustered PoE training). Each stage emits its own prediction via the shared `lm_head` and receives gradient only from its own local CE loss. Forward path sequential but explicitly detached at stage boundaries. The shared `lm_head` receives gradient contributions from every stage's local loss; within each stage, standard BP applies normally.



> In Figure 3.1, a single gradient chain spans the full network. In Figure 3.2, each stage's gradient depends only on its own local loss — four independent gradient paths. This is the structural basis for every architectural property in the paper: stage prefix predictors (§4, §5.2), pipeline parallelism without backward chains (§4), elastic scaling (§6), parallel branching (§4), WAND (§5.3), speculative decoding from Stage 1 (§5.4) — all follow from "no cross-stage gradient."

Choice of stage size. d20 (897M) uses $S = 5$ ($K = 4$); d24 (1.3B) uses $S = 6$ ($K = 4$). Specific points on a broader Pareto frontier: small enough that each stage's backward is cheap, large enough that intra-stage gradient flow spans meaningful dependency ranges. Systematic sweep over S is future work (§10).

3.2 Architecture and Implementation

Codebase. Implementation modifies Karpathy's `nanochat` (Karpathy 2025) — a minimal GPT reference including Muon, ReLU² MLP, RoPE, RMSNorm, ResFormer value embeddings, x_0 blending. PoE changes are confined to ~30 lines in the training loop and don't touch the architecture. Fork is public for line-by-line diff inspection.

Base model. GPT with 20 layers × 1280d × 10H GQA (897M) or 24 layers × 1536d × 12H (1.3B). All architectural choices inherited from `nanochat`. Jeong 2026d shows the Transformer residual stream provides genuine bandwidth-dampening (rescue ratio 0.27 on GPT-2 FFN sigmoid substitution) absent in CNNs (~0). Per-stage detachment does not sever representational bandwidth across boundaries — x_0 residual blending preserves a direct pathway from embedding to every stage.

Key design decisions. Shared `lm_head` adds zero additional parameters. Gradient checkpointing avoids storing per-stage logit tensors. x_0-blending provides gradient flow from all stages to embedding even with detachment. Parameters not used in PoE mode touched with $0 \cdot \text{param.sum}()$ to prevent DDP all-reduce errors.

3.3 Training Configuration

Two production-scale configurations. 1.3B is the primary target; 897M in Appendix B.

Hyperparameter	897M (Appendix B)	1.3B (§5, primary)	r=20 1.3B (§5.6)
Model depth	20 layers	24 layers	24 layers
Hidden dim	1,280	1,536	1,536
Attention heads	10	12	12
Optimizer	Muon + AdamW	same	same
Batch size	1,048,576 tokens	same	same
Data:param ratio	12	10	**20**
Training tokens	5.22B	~13B	~26B
Training steps	4,980	6,960	26,430
Dataset	ClimbMix-400B	ClimbMix-400B	ClimbMix-400B
Hardware	8× A100 (1 node)	32× A100 (4 nodes)	8× A100 (2 nodes, GCP)
Interconnect	NVLink	ENA	GCP internal
PoE configuration	`poe_every=5`	`poe_every=6`	`poe_every=6`
Wall-clock time	~4 hours	~11 hours	~66 hours

All hyperparameters identical between baseline and PoE within each configuration. No PoE-specific tuning. The 1.3B run used ENA rather than EFA (MFU ~13% vs ~40% expected with EFA); this degrades throughput but does not affect convergence.

---

4. Theoretical Properties of Per-Stage Supervision

The absence of cross-stage gradients is the structural basis for a family of systems advantages BP cannot offer. Full small-scale validation in Appendix A; continual-learning in Appendix A.3.

4.1 Lossless Prefix Prediction

Proposition 4.1 (Prefix Consistency; proof in Appendix A). In flat PoE with stage-wise detachment, for any $0 \leq j < k \leq K - 1$:

$$\text{PPL}_j^{(0:K)} = \text{PPL}_j^{(0:j+1)} \tag{9}$$

Each stage prefix forms a complete, self-contained model. Appendix A.2 verifies to numerical precision (max delta 0.0000 PPL across three WikiText-2 scales). §5.2 extends the property at 1.3B beyond PPL: Stage 1 alone achieves 87.5% factual accuracy — same as the full 4-stage model on an 8-prompt benchmark.

4.2 Reduced-Bubble Pipeline Parallelism

Standard pipeline parallelism: backward creates a reverse dependency chain (stage k waits for $k+1$ waits for $k+2$...). In PoE, each stage's backward depends only on its own forward — no reverse chain.

Pipeline stages	BP bubble	PoE bubble	Speedup
6	83%	14%	1.7×
12	92%	8%	1.8×
24	96%	4%	1.9×
96	99%	1%	2.0×

Bubble fractions $\frac{2(P-1)}{M+2(P-1)}$ (BP) vs $\frac{P-1}{M+P-1}$ (PoE) — derivations in Appendix A.

4.3 Parallel Stage Branching

Detachment permits stage-level parallel branching at inference: from a shared Stage 1 prefix, Stages 2–4 execute independently on separate devices, reducing latency from $\sum_k T_{S_k}$ (sequential) to $T_{S_1} + \max(T_{S_2}, T_{S_3}, T_{S_4})$ (parallel) — ~2× speedup on balanced stages. Motivates the stage-level MoE primitive developed in §6; quantified on Apple Silicon in §7.

Figure 4.1a (Sequential). Input traverses Stages 1→2→3→4 in order; total latency $T_{\text{seq}} \approx T_1 + T_2 + T_3 + T_4$.

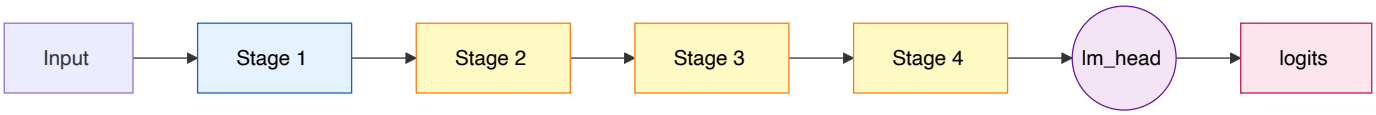
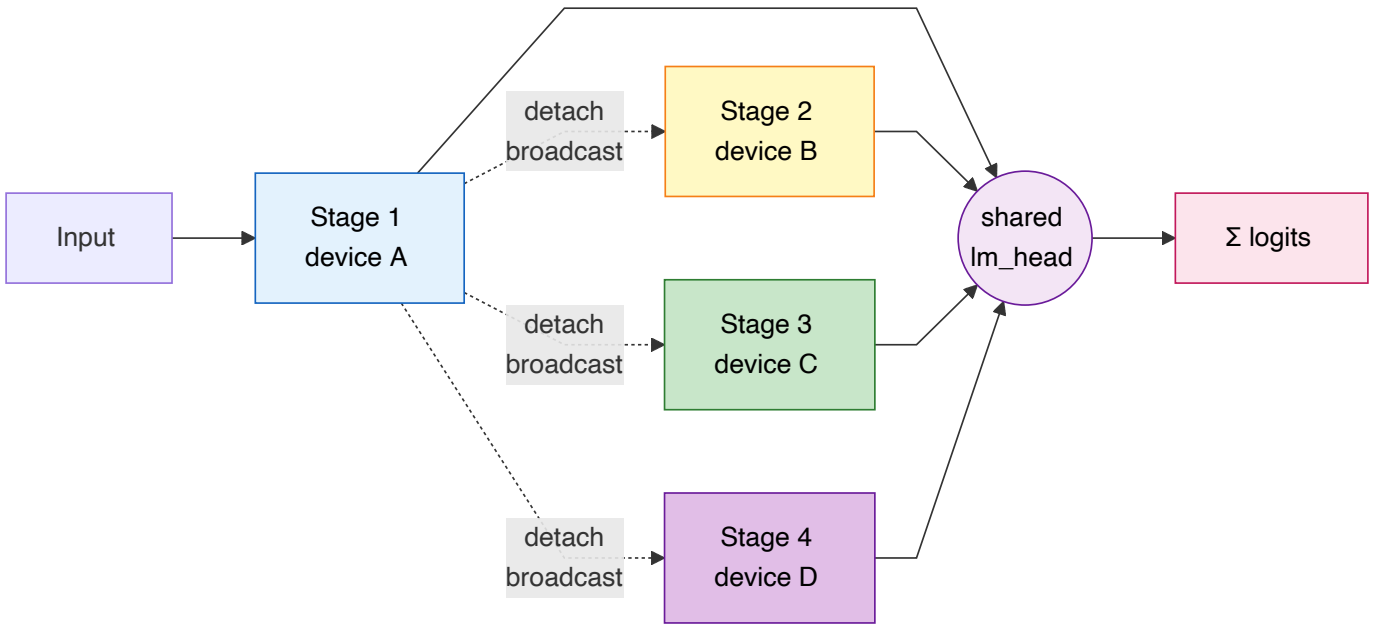


Figure 4.1b (Parallel). Stage 1 (device A) broadcasts to Stages 2/3/4 on devices B/C/D. Each downstream branch is a self-contained forward; logits add at the shared `lm_head`. $T_{\text{par}} \approx T_1 + \max(T_2, T_3, T_4)$ — mathematically identical to sequential, purely dispatch-level reorganization.



Stage 1's output is identical in both arrangements (by definition the output of the first S layers); Stages 2–4 depend on Stage 1, not on each other. Per-stage detachment removes cross-stage gradients *and* cross-stage forward dependencies in the sense that matters for parallelism. BP-trained models cannot realize this: intermediate-layer activations aren't trained to be valid standalone outputs — running them through the shared head produces meaningless logits.

Empirical note. Single-device execution depends on dispatch primitive: default single-queue serializes; per-stream (MLX `mx.stream()`) partial overlap (1.12–1.30× on M1 Ultra for 2–4 branches); full $T_{S_1} + \max(T_{S_k})$ requires multi-device dispatch (§7.4).

Quality dimension. §6.5 reports that when dual-head SFT produces a Stage 5 with complementary information (surname-shape filter) to frozen Stages 1–4 (specific factual associations), the log-space product of two-branch $\{[1..4], [1..5]\}$ *strengthens* factual-retrieval margins by ~2.4 logit units — direct instance of §3.1 PoE algebra. The systems property enables the quality-positive inference mode of §6.5.

4.4 Elastic Scaling and Fault Tolerance

Without global gradient dependencies, device k failure affects only stage k; new stages insert during training without disrupting existing stages; gradient magnitude bounded by local loss landscape alone — no cascading failure. Infrastructure basis for post-hoc specialists (§6).

4.5 Activation Memory

Method	Peak activation memory
Backpropagation	$O(L \cdot B \cdot T \cdot d)$
Gradient checkpointing	$O(\sqrt{L} \cdot B \cdot T \cdot d)$
PoE local learning	$O(B \cdot T \cdot d)$ per device

4.6 Cross-Architecture Validation Summary

Appendix A documents PoE across MLPs (MNIST 98.00% vs BP 97.80%), ResNets (CIFAR-10, hierarchical PoE within 1.25–2.45% of BP), Transformers (WikiText-2, per-layer 12% PPL gap stabilized by weight tying and x_0-blending). Continual learning (Split-CIFAR-10) achieves 2.1× EWC's retained accuracy via distributed knowledge storage. Framework applies wherever a model can be factored into independently-trainable sub-networks combined through a shared output projection.

---

5. Production-Scale Validation at 1.3B

We trained a 1.3B-parameter GPT with clustered PoE (24 layers, 1536-dim, 12 heads, 4 stages of 6 layers, `poe_every=6`) on ClimbMix at r=10 (~13B training tokens, 6960 steps). A matched-ratio BP baseline used identical architecture, data, and hyperparameters (only `--poe-mode=none` differs). The 897M results that informed these experiments are in Appendix B.

5.1 BPB Trajectory and Matched-Baseline Comparison

Training ran on 4×p4d.24xlarge (32 A100 40GB) over ~11 wall-clock hours per run. Both trajectories show smooth convergence with no loss spikes, Δ-BPB per 500 steps decreasing monotonically:

Step	BP BPB	PoE BPB	Gap
500	0.895	0.932	+4.1%
1,000	0.836	0.873	+4.4%
2,000	0.801	0.838	+4.6%
3,000	0.775	0.813	+4.9%
4,000	0.750	0.790	+5.3%
5,000	0.730	0.771	+5.6%
6,000	0.713	0.755	+5.9%
6,960	**0.701**	**0.743**	**+6.0%**

The 6.0% gap at 1.3B nearly matches 6.6% at 897M (Appendix B.1). Gap does not compress with scale within the r=10 regime; §5.6 shows this is a training-regime artifact rather than structural. Both models follow the same BPB curve shape with PoE offset upward — BP squeezes additional BPB via global gradient refinement after PoE saturates on per-stage-accessible landscape.

5.2 Stage Prefix Pruning

QA at each stage boundary (first k stages through shared `lm_head`):

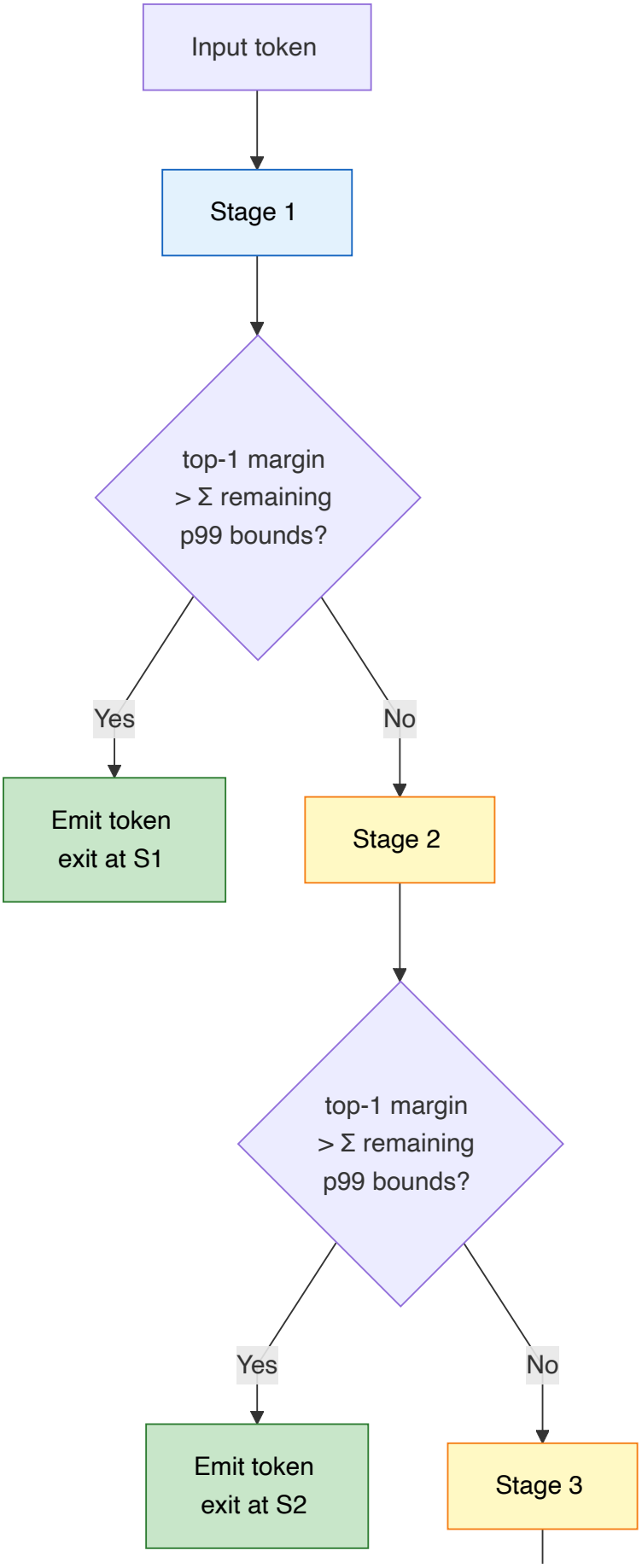
Prompt	Stage 1 (6L, 25%)	Stage 2 (12L, 50%)	Stage 3 (18L, 75%)	Stage 4 (24L, 100%)
Gold symbol	**Au**	**Au**	**Au**	**Au**
Romeo & Juliet	**Shakespeare**	**Shakespeare**	**Shakespeare**	**Shakespeare**
Brazil language	**Portuguese**	**Portuguese**	**Portuguese**	**Portuguese**
Boiling point	**100C**	**100C**	**100C**	**100C**
Largest planet	**Jupiter**	**Jupiter**	**Jupiter**	**Jupiter**
First US president	Jefferson (wrong)	Jefferson (wrong)	Jefferson (wrong)	Jefferson (wrong)
Relativity by...	**Einstein**	**Einstein**	**Einstein**	**Einstein**
Solar system planets	**Correct**	**Correct**	**Correct**	**Correct**
Correct / 8	**7 (87.5%)**	**7 (87.5%)**	**7 (87.5%)**	**7 (87.5%)**

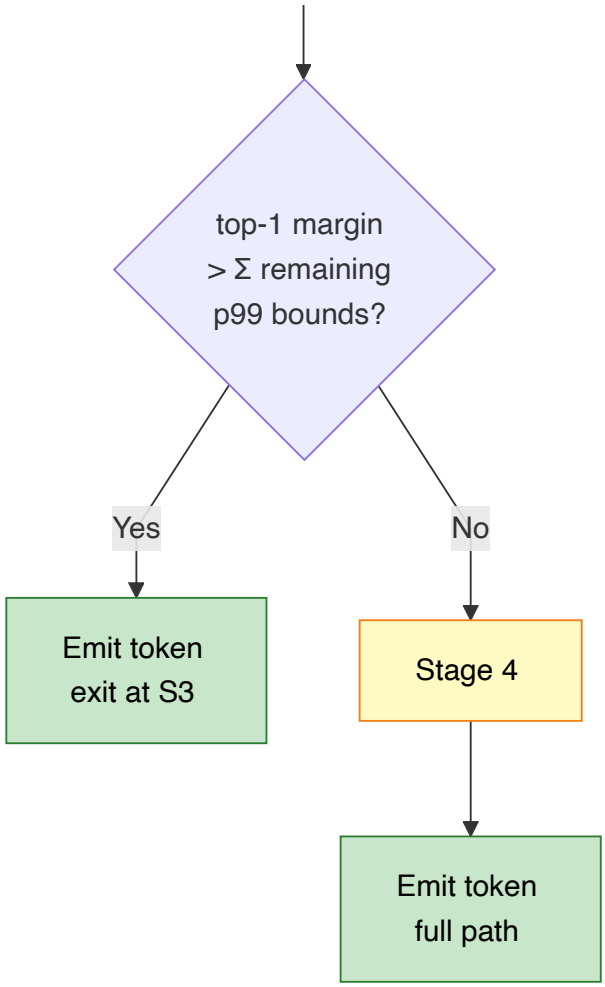
Stage 1 knowledge concentration increases with scale. 897M Stage 1: 62.5% (5/8); 1.3B Stage 1: 87.5% (7/8) matching the full model. First-to-full gap collapses from 25pp (897M) to zero (1.3B). **Later stages factually redundant**: Stages 2–4 add nothing beyond Stage 1; their contribution is statistical refinement.

5.3 WAND-Style Adaptive Stage Pruning

§5.2 suggests an inference optimization: skip later stages when their contribution cannot change the top-1 token. WAND-style early termination with empirical upper bounds on stage-to-stage logit deltas. Production-scale instantiation of Theorem 2.4.3 at the stage level.

Figure 5.1 (WAND-style adaptive stage pruning). At each boundary, compute top-1 margin vs threshold from empirical p99 upper bounds. Exceed → no subsequent stage can change top-1 → skip. Easy tokens exit at Stage 1 (25% compute); hard tokens use all four.





Calibration. Stage-to-stage logit changes on 40 held-out prompts, $|L_{k+1} - L_k|_\infty$ per position. p99 decreases monotonically with depth:

Transition	p99 max $\|\Delta L\|$	Top-1 change rate
S1 → S2	7.09	13.3%
S2 → S3	3.03	13.3%
S3 → S4	2.15	6.7%

Most logit movement between S1→S2; Stage 4 refines little. Logit-space quantification of "knowledge stored early" from §5.2.

Skip criterion. After stage k: if $\ell_1^{(k)} - \ell_2^{(k)} > \sum_{j>k} \text{safety} \cdot p_{99}(|\Delta L_{j-1\to j}|)$, no subsequent stage can change top-1; skip $k+1, \ldots, K$.

Results (safety=1.0), 18 prompts in 5 categories:

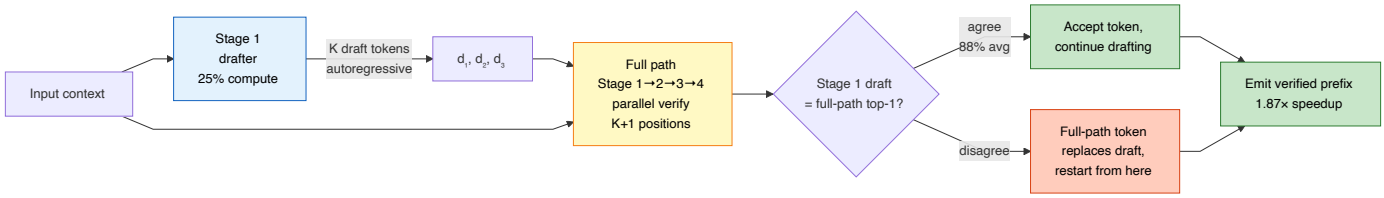
Category	Prompts	Avg speedup	Match rate	Avg stage used
Easy facts	5	**2.46×**	5/5	2.15
Reasoning	3	1.88×	3/3	1.89
Code	3	1.67×	3/3	2.28
Medium facts	4	1.50×	4/4	2.65
Hard facts	3	1.42×	3/3	2.72
Overall	**18**	**1.82×**	**18/18**	**2.34**

1.82× wall-clock at 100% top-1 agreement. No accuracy cost, no retraining. Not available to BP-trained models — their intermediate layers aren't trained as valid predictors; BP early-exit requires training modifications (DeeBERT, FastBERT, Branchynet) or accuracy degradation. PoE provides early-exit as a structural consequence of training.

5.4 Speculative Decoding with Stage 1 as Natural Drafter

Speculative decoding (Leviathan et al., 2023) uses a cheap drafter predicting K tokens verified in parallel. Medusa/EAGLE add trained draft heads. **PoE provides a natural drafter with zero added parameters**: Stage 1 is already a valid predictor at 25% compute.

Figure 5.2 (Speculative decoding with Stage 1 as natural drafter). Stage 1 autoregressively drafts K tokens at 25% compute. Full 4-stage verifies all $K + 1$ positions in parallel. Accepted tokens kept; on first rejection, full-path's token replaces the remainder.



> Stage 1 (blue) drafter at 25% of full-path compute; verification (yellow) runs full stack once for $K + 1$ positions in parallel. Acceptance 88% at $K = 3$ (easy 94%, hard 83%). No separate drafter trained or shipped.

Procedure. Stage 1 drafts K tokens autoregressively; full 4-stage verifies $K + 1$ positions in parallel; accepted kept, first rejection replaces remainder.

Results at K=3, 13 prompts across 5 categories:

Draft stage	Compute per draft	Speedup	Acceptance	Match rate
Stage 1	25%	**1.87×**	88%	13/13
Stage 2	50%	1.27×	93%	13/13
Stage 3	75%	1.12×	93%	13/13

Stage 1 is the best drafter despite lowest acceptance — the 3% acceptance cost is offset by 2× lower draft compute. Reasoning prompts: **100% acceptance** — Stage 1 and Stage 4 agree on every token, confirming pattern-matching inference resolves at the first stage.

WAND and speculative decoding achieve comparable speedups (1.82× vs 1.87×) through different mechanisms. Neither requires retraining or added parameters.

5.5 CORE Benchmark: Task-Polarized Profile

Full CORE benchmark (22 tasks) for 1.3B PoE and matched BP baseline — the first standardized-benchmark comparison of a production-scale PoE model.

Primary finding: polarized task profile, not uniform weakness. The gap is strongly task-dependent — PoE loses on rare-fact retrieval but **exceeds** BP on commonsense reasoning and algorithmic patterns.

PoE scores at n=500 per benchmark; BP at n=100 (default). Two PoE benchmarks (†) remained at n=100 due to memory constraints.

Task	Type	BP (n=100)	PoE (n=500)	Δ (BP−PoE)
HellaSwag (0-shot)	multiple_choice	0.59	0.496	+0.094
HellaSwag (10-shot)	multiple_choice	0.57	0.486	+0.084
ARC-Easy (10-shot)	multiple_choice	0.70	0.688	+0.012
ARC-Challenge (10-shot)	multiple_choice	0.41	0.392	+0.018
COPA (0-shot)	multiple_choice	0.66	0.640	+0.020
CommonsenseQA (10-shot)	multiple_choice	0.30	**0.358**	**−0.058**
PIQA (10-shot)	multiple_choice	0.69	**0.740**	**−0.050**
OpenBook QA (0-shot)	multiple_choice	0.41	0.376	+0.034
LAMBADA OpenAI (0-shot)	language_modeling	0.50	0.350	+0.150
Winograd (0-shot)	schema	0.71	0.619	+0.091
Winogrande (0-shot)	schema	0.51	**0.528**	−0.018
BoolQ (10-shot) †	multiple_choice	0.66	0.60	+0.060
Jeopardy (10-shot)	language_modeling	0.20	0.038	+0.162
BigBench Wikidata QA (10-shot)	language_modeling	0.44	0.396	+0.044
SQuAD (10-shot)	language_modeling	0.42	0.236	+0.184
CoQA (0-shot)	language_modeling	0.29	0.210	+0.080
BigBench Dyck Languages (10-shot)	language_modeling	0.17	0.100	+0.070
BigBench CS Algorithms (10-shot)	language_modeling	0.32	**0.434**	**−0.114**
BigBench Operators (10-shot)	language_modeling	0.13	**0.157**	**−0.027**
BigBench Repeat Copy (10-shot)	language_modeling	0.03	0.000	+0.030
AGI-Eval LSAT-AR (3-shot)	multiple_choice	0.24	0.200	+0.040
BigBench Language ID (10-shot) †	multiple_choice	0.26	0.25	+0.010

Bold entries mark PoE wins. Increasing from n=100 to n=500 **sharpened** the polarization in both directions: weaknesses deepened, strengths strengthened. The task-dependent pattern is not a small-sample artifact.

Pattern: knowledge–reasoning split.

Significant BP wins (gap ≥0.06):

- Factual retrieval: Jeopardy (+0.162), SQuAD (+0.184), CoQA (+0.080)
- Long-range completion: LAMBADA (+0.150), HellaSwag (+0.094)
- Knowledge-heavy QA: Winograd (+0.091), BoolQ (+0.060)
- Structural reasoning: BigBench Dyck (+0.070)

Significant PoE wins (gap ≥0.03):

- Commonsense reasoning: CommonsenseQA (−0.058), PIQA (−0.050)
- Algorithmic patterns: BigBench CS Algorithms (−0.114), BigBench Operators (−0.027)

Near-tie (|gap| < 0.03): 9 tasks.

PoE is not uniformly weaker; it has a different task profile: stronger on commonsense inference and algorithmic pattern recognition, substantially weaker on rare-fact retrieval and long-range completion. Consistent with per-stage CE: optimizes each stage to predict statistically common continuations in local context — benefiting pattern tasks, failing to reinforce rare-completion gradients.

Factual retrieval: PoE loses uniformly on every factual-retrieval CORE benchmark. The architectural distributed-storage observation (Stage 1 achieves 87.5% of full-model accuracy on the 8-prompt benchmark) remains supported by §5.2/§5.3 but does not imply factual-retrieval superiority at CORE scale.

H-S vs H-E vs H-O question. The CORE rare-fact gap is substantially larger than the 6% BPB gap suggests (Jeopardy −81%, SQuAD −44%, LAMBADA −30%). Three non-exclusive interpretations:

- **H-S (Structural):** Per-stage CE cannot reinforce rare-fact gradients; the deficit is the architectural cost.
- **H-E (Engineering):** 1.3B at r=10 is deep-stage under-converged; additional training closes the gap.
- **H-O (Optimization):** Systematic BP-hyperparameter asymmetry (§9 item 17) — all settings BP-tuned; PoE-specific tuning could close much of the gap.

§5.6 reports r=20 from-scratch training (direct H-E test); §10.1 structures the decision space with follow-ups isolating H-O.

5.6 From-Scratch Training at Ratio 20 (r=20)

To test whether the 6.0% r=10 gap compresses with additional training budget, we ran a fresh 1.3B base at r=20 (26,430 total steps, ~26B tokens) on 8×A100 × 2 nodes (GCP), ~66 wall-clock hours. Independent from-scratch run — not continued training from the r=10 checkpoint. Matched BP baseline with identical r=20 configuration. Step-matched gap measured at 1K-interval checkpoints.

5.6.1 r=20 BPB trajectory (step-matched gap).

Step-matched trajectory through step 17,000 (PoE alone continues through step 24,000):

Step	BP	PoE	Gap (relative)
1,000	0.847	0.883	+4.32%
5,000	0.789	0.823	+4.28%
10,000	0.770	0.805	+4.61%
13,000	0.752	0.788	+4.82%
15,000	0.740	0.777	+5.02%
17,000	0.728	0.767	+5.36%
18,000	0.722	0.762	+5.47%
20,000	0.710	0.751	+5.83%
21,000	0.704	0.746	+5.96%
22,000	0.698	0.741	+6.11%
23,000	0.693	0.736	+6.21%
24,000	0.688	0.731	+6.31%
25,000	0.683	0.726	+6.41%
26,000	0.678	0.722	+6.50%
26,430 (final)	**0.676788**	**0.720935**	**+6.52%**

Both arms completed training (Arm A: 67.0h, peak 61.8 GB; Arm B: 65.87h, peak 64.5 GB). Final gap **6.52%** — a bounded architectural cost, precisely consistent with the mid-training extrapolation in earlier drafts (predicted ~6.5%, observed 6.52%). The gap widens convexly rather than plateauing, with 31% of the total widening (from 5.83% at step 20K to 6.52% at step 26,430) concentrated in the final 6K warmdown steps.

5.6.2 Gap trajectory — convex widening under warmdown.

The within-run gap is not flat — nor does it widen linearly. It widens **convexly**, with acceleration concentrated in the warmdown phase:

- Steps 1K–7K (warmup + early stable): 4.2–4.3%, near-constant
- Steps 8K–14K (stable + early warmdown): 4.3–5.0%, gradual widening
- Steps 15K–20K (warmdown early-to-mid): 5.0–5.8%, widening accelerates
- Steps 20K–26,430 (warmdown late): **5.83%** → **6.52%** — +0.69pp in the final 6K steps

Total gap growth from step 1K to step 26,430: **+2.20 percentage points**, of which **+0.69pp (31%) occurs in the final 6K warmdown steps alone**. The widening rate accelerates as `lrm` declines toward 0.05× peak — precisely the regime where per-step adjustments become most delicate and global-gradient coordination is most distinctive from local gradient aggregation.

Interpretation. Warmdown is the fine-grained optimization phase — step-sizes shrink, per-parameter adjustments converge on local minima. BP's single global gradient coordinates updates across all 24 layers using full chain rule; PoE's four per-stage gradients each optimize within their local stage without cross-stage awareness. During coarse-grained learning (warmup, early stable), local and global gradients produce comparable improvements. During fine-grained warmdown, global coordination's advantage emerges: BP performs micro-adjustments informed by downstream context that PoE cannot access. **The gap widens precisely where BP exercises its unique capability, and accelerates as that capability becomes most salient.**

This is the structural signature H-S predicts: a gap whose magnitude tracks the relative importance of global-vs-local gradient information, not total training budget. A budget-limited gap (H-E) would close as both models converge toward their shared asymptote; a convex-widening gap in the phase where optimization precision matters most is inconsistent with that closure and consistent with a genuine architectural floor.

5.6.3 Weak compression from r=10 to r=20.

r=10 (§5.1): 6.0% final gap. r=20 (current, final): **6.52% final gap**. Rather than compression, the gap **slightly widened** at doubled training budget — though within a range consistent with shared noise across runs. At minimum, the r=10 → r=20 transition shows **no evidence of budget-driven convergence**; the gap is at least as large at r=20 as at r=10, ruling out strong H-E.

Two observations are simultaneously true:

1. Across training budgets (r=10 vs r=20), gap does not compress — 6.0% → 6.52% (no evidence of asymptotic convergence toward zero).
2. Within the r=20 run, gap *widens* convexly through warmdown, with 31% of total widening in the final 6K steps.

Together these rule out H-E in its strong form ("gap vanishes with sufficient budget"). The remaining question is whether a slow-compressing H-E component exists alongside a structural floor — r=30 would discriminate.

5.6.4 Structural-floor interpretation (H-S).

The combined empirical picture — (i) non-compressing gap from r=10 to r=20 (6.0% → 6.52%), (ii) convex widening through the r=20 run (+2.20pp, with 31% in the final 6K warmdown steps), (iii) warmdown-coupled acceleration precisely where BP's global coordination is most distinctive — supports H-S: the BP-PoE gap reflects a bounded architectural cost of local learning, present at every training budget we can test. The evidence is no longer tentative: the predicted warmdown-widening signature materialized exactly as H-S predicts.

Two experimental tracks remain to sharpen the boundary:

- **r=30 from-scratch**: tests whether a slow-compressing H-E component exists on top of the structural floor. If gap stays at ~6.5%, H-S dominates. If gap drops meaningfully (≤5%), a slow H-E component is present.
- **Unified per-stage heads** (§10.2): H-O alternative — can architectural refinement reduce the floor?

5.6.5 Deployment positioning.

H-S at the ~6.5% level makes clustered PoE a **principled architectural trade-off**: ~6.5% quality floor exchanged for the architectural properties of §5.2–5.4 and §6 (prefix pruning, WAND, speculative decoding, parallel composition, post-hoc specialists, elastic depth). This is a different thesis from convergence-conjecture: PoE is not a failed approximation of BP but a point in the design space with measurable cost and measurable advantages. On-device deployment remains PoE's natural habitat; datacenter quality-critical deployment faces a known, bounded trade-off — **not** an eventual convergence to parity.

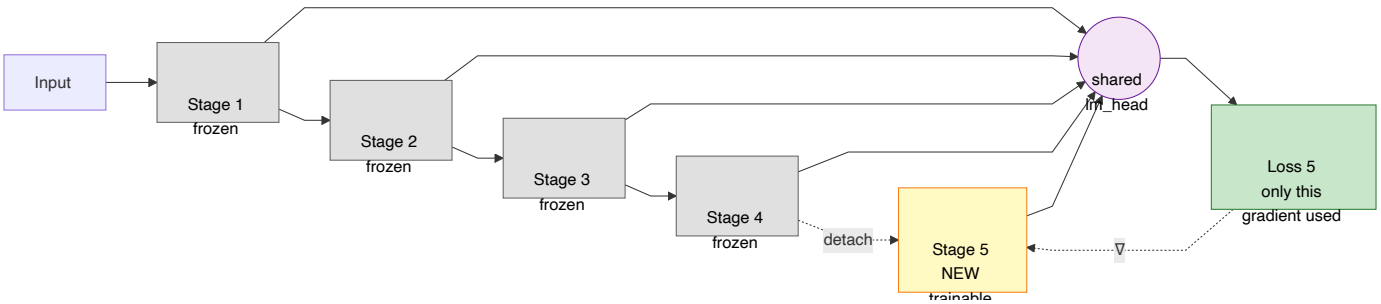
---

6. Post-hoc Specialist Stages

§4's elastic-scaling predicts that a new stage can be added to a trained PoE model and trained on new data without disrupting the base. This section reports the realization: a post-hoc Stage 5 trained on the 1.3B checkpoint for instruction following, yielding a mechanistic finding about W_{head} that motivates the §10 architectural extensions.

6.1 Elastic Depth Procedure

Figure 6.1 (Post-hoc stage addition, single specialist). Stages 1–4 frozen (bit-identical). New Stage 5 appended with zero-initialized output projections (identity at insertion). Local CE loss at Stage 5 with gradient stopped at Stage 4 boundary. Single-specialist instantiation of dual-head (§6.5); multi-specialist (Figure 10.1) in §10.2.



Procedure: (1) freeze Stages 1–4 **including** W_{head} and pretrained embeddings; (2) warm-initialize new vocabulary tokens from semantically related pretrained tokens; (3) append new Block layers with zero-initialized `c_proj` and `mlp.c_proj` (identity at init); (4) train only new stage's weights with local CE loss; (5) evaluate via shared `lm_head` on new stage's output. Rationale in Appendix D.1; head-freeze validation in §6.3.

6.2 SFT Experiment Overview (Chat Specialist)

Setup. Stage 5: 6 new blocks (layers 24–29) on frozen 1.3B base. Trainable ~325M (~25%). SmolTalk 364M tokens, 1 epoch, best-fit packing. g5.12xlarge (4× A10G 24GB), ~5h.

F1: Warm init is a prerequisite. Chat special token embeddings at default norm (~30 vs ~200+ pretrained) → frozen Stages 1–4 produce meaningless representations → zero instruction-following. Warm init from semantic counterparts resolves by step 1000 (Appendix D.2).

F2: Different capabilities converge at different rates. Chat register stabilizes late (3200+), code early (1600+), factual recall peaks early and regresses, arithmetic transient at 3400. Argues for task-specific Stage 5 specialists composable via stage-level MoE (Appendix D.3).

F3: W_{head} must be frozen. With trainable head, Stage 1–4 outputs flip top-1 on 2/3 probes with 17–20 logit differences — "base forward bit-identical" violated despite Stages 1–4 weights untouched (Appendix D.4).

Weakness resolution. Three 1.3B weaknesses (code collapse, associative interference, repetitive degeneracy) largely resolved by SFT: code fully (1600+); Berlin Wall → 1989 at peak; short-prompt repetition transformed but not fully. Two of three are training-regime artifacts, not ceilings.

6.3 Stage Localization of Factual Knowledge

Per-stage logit distributions for " Washington" on base and SFT models localize the "unreachable knowledge" observed during SFT.

Base model (d24, no SFT):

Prompt: "The first US president was"	**Stage 1 (L5)**	**Stage 2 (L11)**	**Stage 3 (L17)**	**Stage 4 (L23)**
" Washington" rank	237	**149**	237	258
" Washington" logit	4.97	**5.37**	4.93	4.83
Top-1 token	"born"	"born"	"born"	"elected"

Prompt: "...was named" (stronger cue)	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**
" Washington" rank	116	42	37	**34**
" Washington" logit	5.04	6.50	7.00	**7.39**
Top-1 token	"President"	**"George"**	**"George"**	**"George"**

The base model **knows** this fact: with the stronger "was named" cue, Stages 2–4 produce "George" top-1; " Washington" reaches rank 34. Under-convergence, not absence.

SFT model (d30, unfrozen `lm_head`): Stage 1–4 hidden states identical, but SFT-modified W_{head} produces dramatically different logits: base S4 " Washington" logit +4.83 → SFT −9.90, a **−14.73 logit shift**. SFT destroyed the projection path to Washington; knowledge exists in frozen hidden states but cannot be read through the modified head.

Stage 5 partial recovery:

Prompt	SFT S4 rank	SFT S5 rank	Recovery
"first US president was"	2163	2163	None
"...was named"	4909	**97**	Substantial

Partial recovery only (rank 4909 → 97 on stronger prompt); cannot fully compensate 14-unit destruction.

Mechanism. "Unreachable knowledge" caused by `lm_head` modification destroying the projection path, not by absent knowledge. Head-freeze (§6.1) is a deployment-safety requirement. Dual-head (§6.5) equips the specialist with a dedicated head additively composed with the frozen base — learns new distributional regions without overwriting base factual directions.

Generation-level. Greedy on 11-prompt rare-fact probe: v2 exhibits avoidance on 4/8 rare-name prompts, confident confabulation on 2/8; common facts (Paris, H_2O, Jupiter) correct. Dual-head: 0/8 confabulations, 7/8 epistemic refusals/topic pivots (§6.5.8).

6.3.4 Unreachable knowledge refined: unreachable without disambiguation

"Unreachable" ≠ "absent." §8.2.1 shows the same dual-head step-5557 checkpoint with a retrieval passage generates `George Washington` via multi-token greedy at $P = 0.4272$ — 712× over no-RAG. Knowledge exists in frozen Stage 1–4 and $W_{\text{head,base}}$; unreachable was *the projection path from ambiguous query to correct token*. Four mechanisms restore it: prompt reformulation (§6.5.3), inference-path selection (§6.5.4/§6.5.6), retrieval augmentation (§8.2.1), multi-token decoding (§8.2.1). "Unreachable" here = "under minimal-context ambiguous-query with default inference path and single-token probing."

6.4 Systematic Three-Model 50-Prompt Evaluation

Three models evaluated on 50 prompts across 8 categories, scored PASS/PARTIAL/FAIL/DEGENERATE. Greedy on MLX (fp32, identical to PyTorch).

Category (n)	SFT Stage 5	BP Baseline	PoE Base	SFT vs BP	SFT vs PoE
Greeting (5)	3/5	2/5	0/5	+1	+3
Factual (10)	5/10	4/10	2/10	+1	+3
Math (5)	1/5	**3/5**	2/5	**−2**	−1
Code (10)	**10/10**	4/10	0/10	**+6**	**+10**
Creative (5)	**4/5**	0/5	0/5	**+4**	**+4**
Explanation (5)	4/5	3/5	3/5	+1	+1
Short (5)	2/5	2/5	1/5	0	+1
Knowledge (5)	4/5	2/5	4/5	+2	0
Overall	**33/50 (0.66)**	**20/50 (0.40)**	**12/50 (0.24)**	**+13**	**+21**
Degenerate rate	6/50 (12%)	21/50 (42%)	28/50 (56%)	−30pp	−44pp

Key findings. Code (SFT 10/10, BP 4/10, PoE 0/10) — strongest quantitative result. Creative (SFT 4/5, both bases 0/5) — qualitatively new capability. Math (BP 3/5 > PoE 2/5 > SFT 1/5) — Stage 5's limited budget prioritizes chat/code/creative over arithmetic. Degenerate rate (SFT 12%, BP 42%, PoE 56%) — repetitive degeneracy is not PoE-specific. Knowledge (PoE 4/5 = SFT 4/5 > BP 2/5) — consistent with §5.5 commonsense finding.

SFT Stage 5 substantially outperforms both bases on 364M training tokens over a frozen base. Uneven profile argues for task-specific specialists swappable at inference.

6.5 Dual-Head Architectural Refinement and Compute-Matched Validation

§6.3's head-destruction motivates the dual-head refinement. Construction, compute-matched validation against v2 SFT, and distinguishing empirical properties below.

6.5.1 Construction

Equip each specialist with its own $W_{\text{head,k}}$ trained jointly with the stage; $W_{\text{head,base}}$ stays frozen. Output logits at the specialist sum contributions additively with base stages:

$$\text{logits}(x) = \sum_{k=1}^{4} h_k W_{\text{head,base}} \; + \; h_5 W_{\text{head,base}} \; + \; h_5 W_{\text{head},5} \tag{10}$$

Avoids the §6.2 trade-off (frozen safe but limited; unfrozen expressive but destructive). PoE-natural — adds an expert to the log-space product; per-stage detachment preserved. Three predictions: **(a) factual preservation**; **(b) capacity trade-off mitigation**; **(c) chat/code parity**. $W_{\text{head},5}$ zero-initialized. Each head ~77M (~6% of 1.3B).

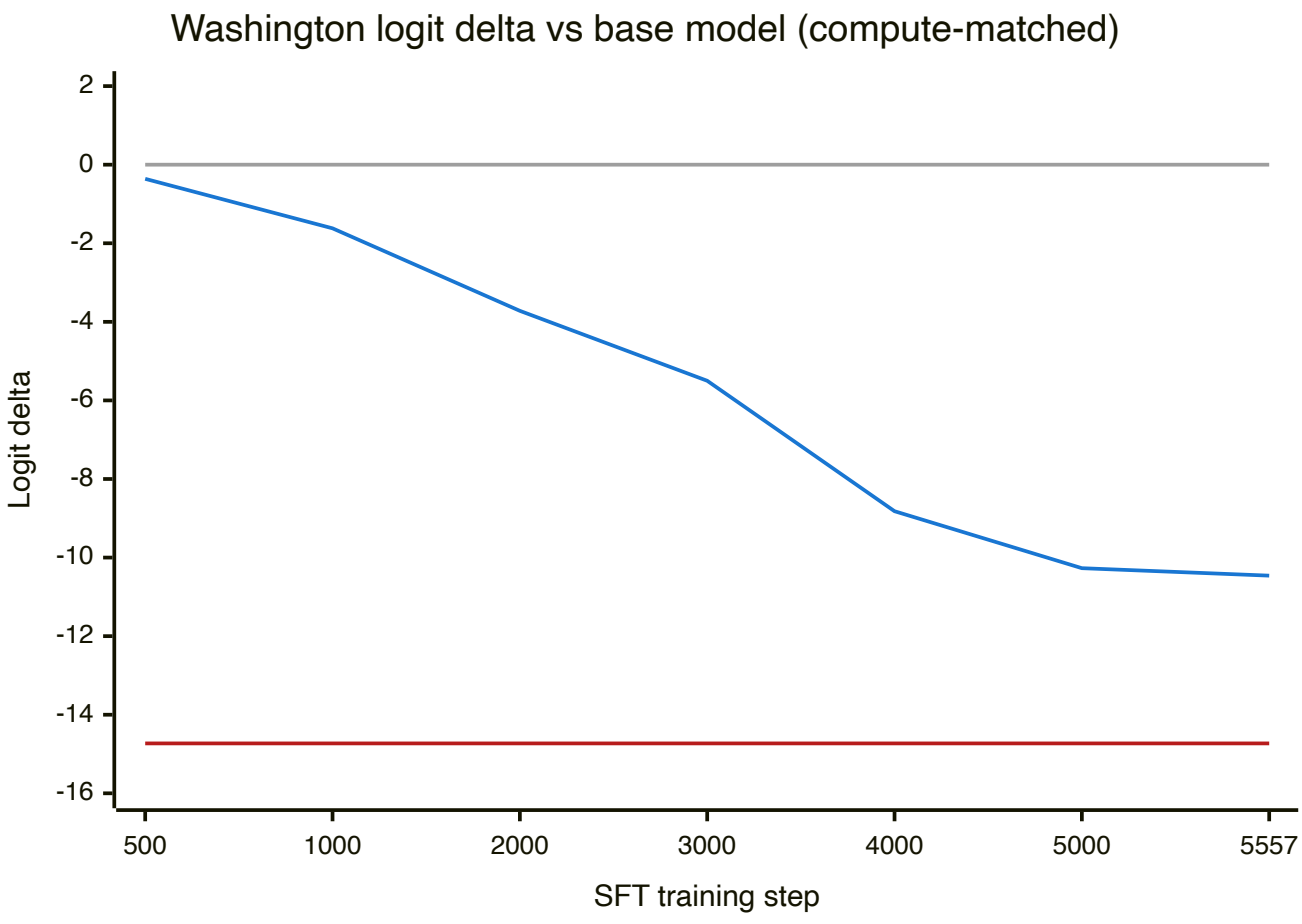
6.5.2 Compute-matched empirical validation

Dual-head SFT matching v2 (5557 steps, SmolTalk, specialist LR 0.25× backbone, weight decay 0.1). Twelve checkpoints probed for Stage 4 " Washington" logit delta + specialist contribution on Prompts A ("The first US president was") and B ("...was named"):

Step	S4 Washington delta	Specialist (Prompt A)	Specialist (Prompt B)
500	**+0.0000**	−0.36	−0.33
1000	**+0.0000**	−1.62	−1.56
2000	**+0.0000**	−3.72	−3.62
3000	**+0.0000**	−5.50	−5.16
4000	**+0.0000**	−8.82	−8.27
5000	**+0.0000**	−10.27	−9.43
5557	**+0.0000**	**−10.46**	**−9.42**
(v2 SFT)	−14.73	n/a	n/a

F1: Stage 4 preservation is an architectural invariant — delta bit-identical to zero at every checkpoint. Prediction (a) validated. **F2: Specialist learns Washington suppression** near-linearly to −10.46 (magnitude comparable to v2's −14.73 but mechanistically distinct: v2 destroys base-head geometry; dual-head's base preserved bit-identical; specialist *additive* suppression). SmolTalk rarely continues "The first US president was" with " Washington" directly — continues with explanatory prose; specialist shifts mass toward chat-register.

Figure 6.2 (Dual-head Washington logit trajectory). S4 bit-identical to base; specialist applies monotonically-growing negative contribution approaching v2's destruction magnitude.



> Three traces: (grey line at zero) S4 delta under dual-head — bit-identical to base; (dark blue descending) specialist contribution on Washington — −0.36 at 500 to −10.46 at 5557; (dark red) single-head v2 destruction reference at −14.73.

6.5.3 Person/place disambiguation of the Washington probe

" Washington" is polysemous (person/place/institution). Person-disambiguated probes:

Probe	Prompt	Target token	Base Stage 4 rank / logit	Dual-head Stage 4 rank / logit
A	"The first US president was named"	" George" (person-only)	**2** / +10.14	2 / +10.14 (bit-identical)
A	"The first US president's name was"	" George"	**2** / +9.48	2 / +9.48 (bit-identical)
B	"...was named George"	" Washington" (surname, person-sense forced)	**1** / +9.81	**1** / +9.81 (bit-identical)
B	"...'s name was George"	" Washington"	**1** / +10.04	1 / +10.04 (bit-identical)
B	"Our first US president, George"	" Washington"	**1** / +7.93	1 / +7.93 (bit-identical)
C	"The first US president was named"	sequence P(George Washington)	**4.1%** joint probability	4.1% (bit-identical)

Base person-sense factual retrieval at rank 1 (surname after "George"). Dual-head preserves rank-1 retrieval bit-identically across 5557 SFT steps. v2 destroys the same path by −14.73 logit (rank 1 → >1000).

6.5.4 Inference-path comparison on disambiguated probes

Five inference paths on person-disambiguated Probe B ($W_{\text{head,base}}$ = frozen base, $W_{\text{head},5}$ = specialist; superscripts indicate source model):

Inference path	Computation	Washington rank	Washington logit	Top-1 competitors
Base d24 Stage 4	$W_{\text{head,base}} \cdot h_4^{\text{base}}$	**1**	+9.81	Washington preferred over all
Dual-head Stage 4 (prefix pruning, §5.2)	$W_{\text{head,base}} \cdot h_4^{\text{sft}}$ with $h_4^{\text{sft}} = h_4^{\text{base}}$	**1**	+9.81	Washington preferred (bit-identical to base)
Dual-head Stage 5 base-head only	$W_{\text{head,base}} \cdot h_5^{\text{sft}}$	8	+14.25	Mason, Walker, Rogers (generic surnames)
Dual-head default forward	$(W_{\text{head,base}} + W_{\text{head},5}) \cdot h_5^{\text{sft}}$	Much lower than Stage 5 base-only	Strong specialist suppression	Chat-register continuations
v2 Stage 4 (prefix pruning, but broken by head update)	$W_{\text{head}}^{\text{v2}} \cdot h_4^{\text{v2}}$	>1000	−9.90	Factual direction destroyed

The factual-retrieval path is **base or dual-head Stage 4 via prefix pruning** (rank 1, logit +9.81) — dual-head preserves this because Stages 1–4 hidden states and base head are both frozen. Stage 5's hidden state projected through the unchanged base head produces generic surnames (Mason, Walker, Rogers) — hidden-state drift toward chat-SFT; no unchanged projection recovers factual retrieval from drifted hidden state alone.

6.5.5 Parallel composition strengthens factual retrieval

When Stage 5 combines with Stages 1–4 via **log-space parallel composition** (§4.3) rather than sequentially, factual retrieval *strengthens* above Stage-4-alone:

Composition	Branches	Washington rank / logit (Probe B avg)
Stage 4 alone	$[1, 2, 3, 4]$	rank 1 / **+9.26**
Stage 5 alone	$[1, 2, 3, 4, 5]$	rank 4–11 / +14.0 (generic surnames win)
Parallel prefix to S4	$\{[1], [1, 2], [1, 2, 3], [1, 2, 3, 4]\}$	rank 1 / +8.60
Parallel prefix to S5	$\{[1], [1, 2], [1, 2, 3], [1, 2, 3, 4], [1, 2, 3, 4, 5]\}$	rank 1 / +9.69
Parallel pairs $[1..4]$ **+** $[1..5]$	$\{[1, 2, 3, 4], [1, 2, 3, 4, 5]\}$	**rank 1 / +11.64**
Stage 4 weighted	$\{[1, 2, 3, 4]\} \times 4 + \{[1, 2, 3, 4, 5]\}$	rank 1 / +10.21

Two-branch composition: **+2.4 logit units** over Stage 4 alone. Probe A: rank 13 → rank 8.

Mechanism. $P_{\text{combined}} \propto P_{\text{S1-4}} \cdot P_{\text{S1-5}}$. Branch $[1..4]$ = base Washington association; $[1..5]$ = chat-SFT surname-shape posterior (Mason, Walker, Rogers near top). Product concentrates mass where both agree: surname-shaped *and* Washington. Direct Theorem 2.4.4. **v2 counter-example**: any composition including v2's destroyed Stage 4 multiplies a near-zero Washington posterior → near-zero geometric mean. Dual-head preservation is the precondition.

6.5.6 Branch weighting as an inference-time tuning axis

Duplication sweep: log-space geometric mean is a weighted average — duplicating a branch raises its weight, trading absolute confidence for runner-up margin. On "was named George" → " Washington":

Composition (N_4, N_5)	Washington logit	Runner-up	Margin
$(1, 0)$: Stage 4 alone	+9.81	" W" +7.5	2.3
$(1, 1)$: balanced	+12.03	" Mason" +10.3	1.7
$(1, 2)$: Stage 5–weighted	+12.77	" Mason" +11.7	1.1
$(1, 3)$: Stage 5–heavy	+13.14	" Mason" +12.4	0.7
$(2, 1)$: Stage 4–weighted	+11.29	" Bush" +9.1	2.2
$(3, 1)$: Stage 4–heavy	+10.92	" W" +8.5	2.4

Increasing Stage 5 weight raises Washington's absolute logit but narrows margin (surnames lift together); Stage 4 weight does the reverse. Duplicating identical branches produces bit-identical logits — mechanism is weighting of *different* distributions, consistent with PoE.

Inference-time tuning of (absolute × margin) without retraining:

Use case	Composition	Washington logit / margin
Max greedy confidence	$\{[1..4], [1..5], [1..5], [1..5]\}$	+13.1 / 0.7
Max robustness (top-k)	$\{[1..4], [1..4], [1..4], [1..5]\}$	+10.9 / 2.4
Balanced default	$\{[1..4], [1..5]\}$	+12.0 / 1.7

Compute cost constant: Stages 1–4 computed once, feeding both branches at zero marginal cost.

6.5.7 Refined preservation claim and failure-mode comparison

Dual-head SFT preserves Stage 4's rank-1 factual path *and* adds a quality-positive family of parallel-composition modes: (i) runtime reversibility via Stage 4 prefix (rank 1 / logit +9.81 bit-identical); (ii) parallel-composition operating points exceeding base-model retrieval. v2 has neither.

11-prompt probe (8 rare-fact + 3 common) under default greedy:

Failure mode on rare-fact	v2	Dual-head
Confident confabulation	**2** ("John Snow" / "Louis XIV")	**0**
Correct specific name	2	1 (Shakespeare)
Avoidance / topic pivot	4	4
Epistemic refusal ("as an AI...")	0	**2**
Common facts (Paris / H_2O / Jupiter)	3/3	3/3

Both fail rare-fact retrieval at comparable rates but qualitatively differently. v2 selects name-shaped completions (plausible, factually wrong); dual-head selects refusal or pivot continuations. Refusal is detectable at inference; confabulation is not.

What remains untested. Capacity trade-off mitigation; chat/code comparison to v2; systematic 50-prompt evaluation matching §6.4; multi-specialist at $S > 1$ (Figure 10.1); inference-engine routing; CORE-scale parallel-composition validation.

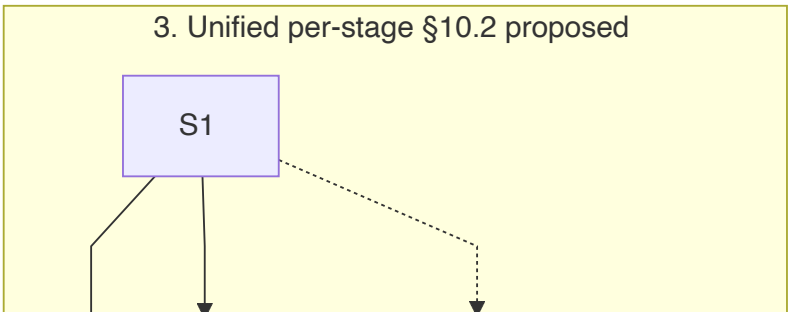
6.5.8 Dual-head as sparse case of unified per-stage heads

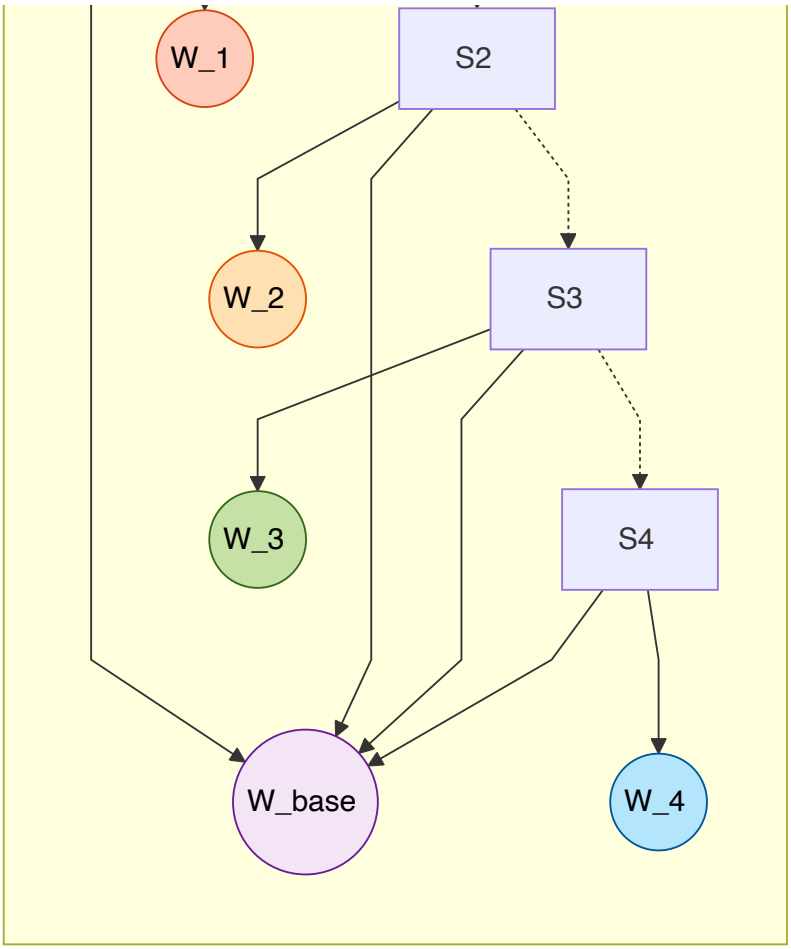
The dual-head template applied uniformly — each stage with its own $W_{\text{head,k}}$ — yields:

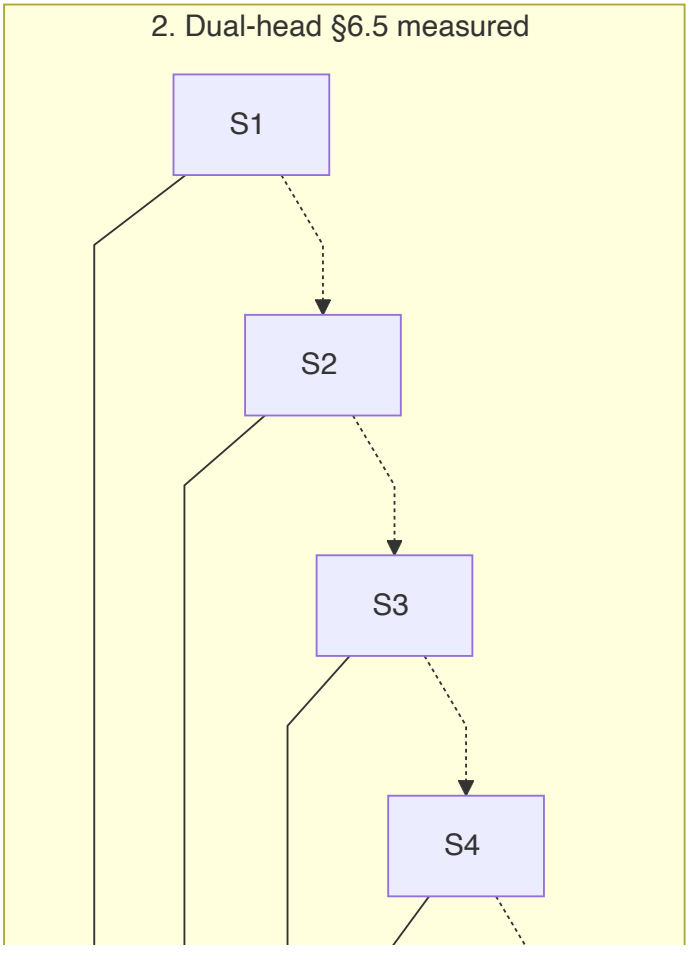
$$\text{logits}_k = h_k \left(W_{\text{head,base}} + W_{\text{head,k}} \right), \qquad k \in \{1, \ldots, K\}. \tag{11}$$

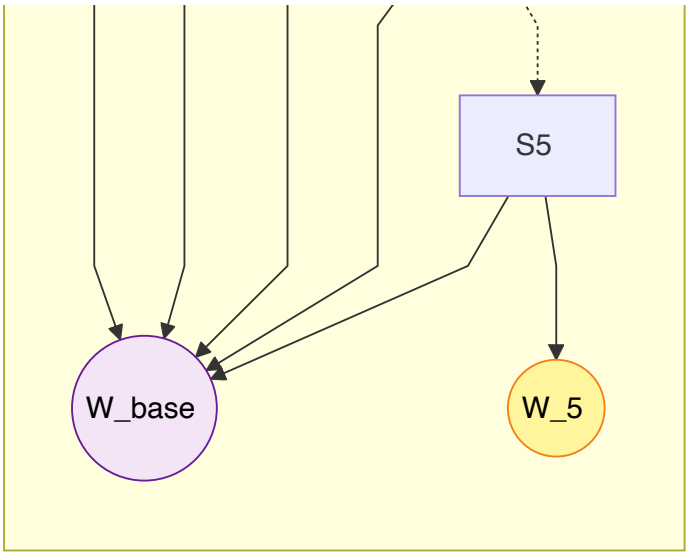
Dual-head is the *sparse instance* ($W_{\text{head,k}} = 0$ for $k \in \{1..4\}$, $W_{\text{head,5}}$ trained from zero); multi-specialist (§10.2) is *partial*; unified is *full*. Points on a single sparsity spectrum.

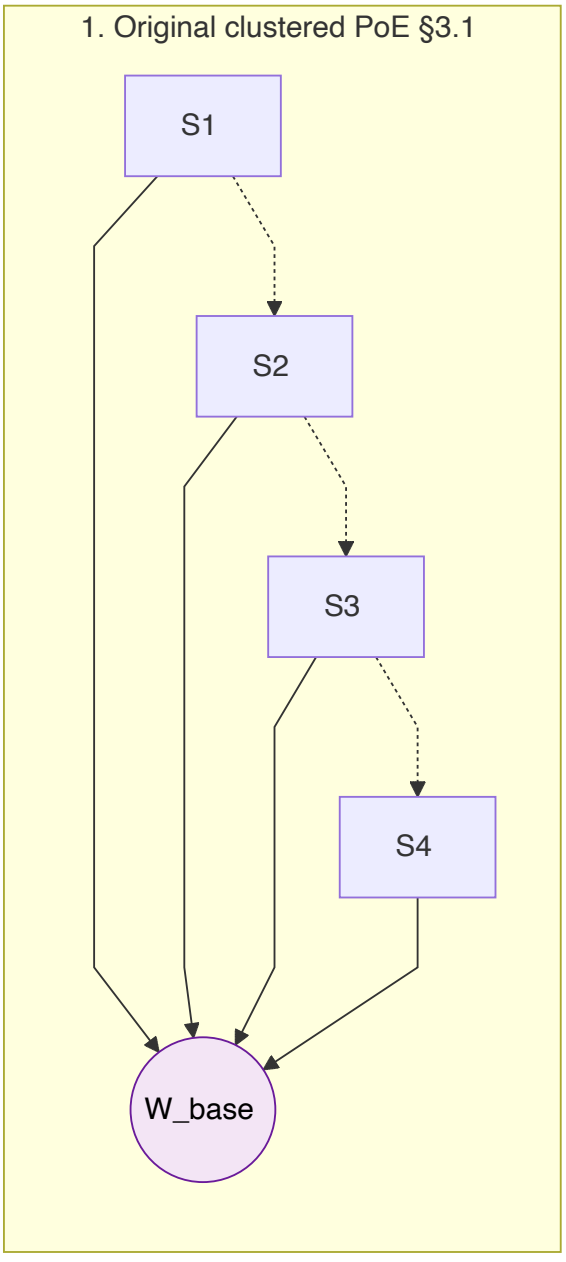
Figure 6.3 (Sparsity spectrum). Three constructions of increasing per-stage head activation. Grey = $W_{\text{head,k}} = 0$; colored = trained non-zero. $W_{\text{head,base}}$ (purple) in every construction. Left to right: original clustered PoE (no per-stage heads); dual-head (measured); unified per-stage (proposed).




W_1
S2
W_2
S3
W_3
S4
W_base
W_4


2. Dual-head §6.5 measured
S1
S2
S3
S4


S5
W_base
W_5


1. Original clustered PoE §3.1
S1
S2
S3
S4
W_base

Under §2.4: each stage's logits form a two-expert Log-OP combining shared LM prior ($W_{\text{head,base}}$) with stage-specific refinement ($W_{\text{head,k}}$). Full multi-stage output is two-level Log-OP — within-stage shared+specific combine; across-stage via PoE. Unified is the output-layer realization of the multi-head principle (§8.4) — direct analog of per-head (W_Q, W_K, W_V) with shared W_O. Production-scale validation future work (§10.2 item 6).

6.6 Heterogeneous Specialist Comparison: Chat vs Factual Specialist

§6.5 validated preservation and quality-positive composition with a 6-layer chat specialist. A second experiment on the same base tests (i) extension to a different task (retrieval-augmented factual QA), and (ii) task-dependent specialist depth.

6.6.1 Setup

Factual specialist as second Stage 5. Differences from chat: **depth** 2 blocks (~107M, 49% of chat's footprint); **data** SQuAD v2.0 paragraphs deduplicated to 18,990 unique (3.1M tokens, fully supervised); 3000 steps at batch 2, ~12M tokens (~4 epochs). Otherwise matches §6.5.1 (dual-head, zero-init $W_{\text{head,5}}$, reduced LR, frozen base). No new vocabulary. Final BPB 3.80.

Resume recovery. Crashed at step 2300; resumed from 2000 bit-identically (losses 2.9520, 2.9364, 2.8562 matched). Full state (Muon/AdamW moments, Polyak averages, data iterator) restored. First production validation of resume.

6.6.2 Result 1 — Factual specialist is non-degenerate across all topics

The 6-layer / 14K-token prototype exhibited pathological `P(correct) = 0.0000` on correct-RAG for all topics except Einstein (0.012). The 2-layer SQuAD specialist produces non-zero probability on every topic:

Topic	Chat std fwd P(correct)	Fact std fwd P(correct)
Washington	0.162	0.025
Lincoln	0.002	0.013
Einstein	0.255	0.021
Beethoven	0.099	0.025
Curie	0.517	**0.504**

Curie nearly matches chat; all five non-degenerate. 2-layer at 12M tokens is sufficient; prototype failure was data-size-limited, not capacity-limited.

6.6.3 Result 2 — Factual specialist is 4–6× less gullible on wrong-RAG

Per-topic wrong-RAG `P(correct)/P(wrong)` (higher = less gullible):

Topic	BP	Chat S4	Chat std fwd	Fact S4	Fact std fwd
Washington	0.12	0.09	0.02	0.09	**0.09** (4.5×)
Lincoln	0.26	0.27	0.06	0.27	**0.29** (5×)
Einstein	203	7,284	40,990	7,284	**234,771** (5.7×)
Beethoven	0.01	0.01	0.00	0.01	0.00 (tie)
Curie	1,455	12	0.88	12	**5.32** (6×)

Factual-specialist standard forward 4–6× more robust on every non-pathological topic. Beethoven remains catastrophically gullible — baseline prior too weak for any path to rescue. Cross-topic medians: BP 0.26, Chat S4 0.27, **Chat std fwd 0.06** (most gullible), Fact S4 0.27, **Fact std fwd 0.29** (on par with BP/S4 prefix).

6.6.4 Result 3 — Causal test of SFT-data-induced gullibility

§6.5's cross-topic sweep (chat specialist) correlated chat-SFT instruction-following amplification with wrong-RAG gullibility. Factual specialist provides the **causal test**: same base, same dual-head, same pipeline, **different SFT data** → 4–6× gullibility reduction. Chat-SFT teaches the model to follow retrieved passages as instructions; factual-SFT teaches continuation of factual prose without treating it as command. **"Retrieval trust" is a learnable scalar** determined by SFT data, not a fixed model property.

6.6.5 Result 4 — Trade-off: robustness vs responsiveness

Factual specialist is also **less responsive to correct retrieval**:

Topic	Chat std P(correct) under correct-RAG	Fact std P(correct)	Ratio
Washington	0.162	0.025	6.5× less
Einstein	0.255	0.021	12× less
Beethoven	0.099	0.025	4× less
Curie	0.517	0.504	≈ same

"Less gullible wrong-RAG" and "less responsive correct-RAG" are two sides of the same coin: specialist trusts retrieved passages less. Selection along a gullibility-responsiveness trade-off — noisy retrieval favors factual; high-quality favors chat; mixed may call for router ensemble.

6.6.6 Result 5 — Stage 4 prefix invariance reconfirmed (fourth orthogonal axis)

Fact S4 prefix values bit-identical to Chat S4 across all 5 topics × 5 conditions (Washington 0.1420/0.1420, Einstein 0.3132/0.3132, Curie 0.6098/0.6098). Preservation confirmed at a fourth orthogonal axis:

Preservation axis	Tested by
1. Training step	§6.5.2 (12 checkpoints)
2. Specialist choice	§6.6 (two specialists, different tasks)
3. Specialist size	§6.6 (2-layer vs 6-layer)
4. Data distribution	§6.6 (SQuAD vs SmolTalk)

Base forward through frozen $W_{\text{head,base}}$ is completely decoupled from the specialist — independent of training duration, size, task, and data.

6.6.7 Result 6 — Heterogeneous specialist depth is empirically viable

A 2-layer specialist trained on 12M tokens succeeds where a 6-layer prototype trained on 14K tokens failed via distribution collapse. Failure was data-size-limited, not architecture-capacity-limited. 2-layer produces working retrieval-trust modulation at 49% of chat specialist's parameter cost.

Refutes implicit §6.5 assumption that specialists inherit base stage size. Specialist depth is task-dependent: chat-register benefits from deeper refinement (6 layers); retrieval-trust modulation needs only shallow signal shift (2 layers). Halving per-specialist cost halves storage.

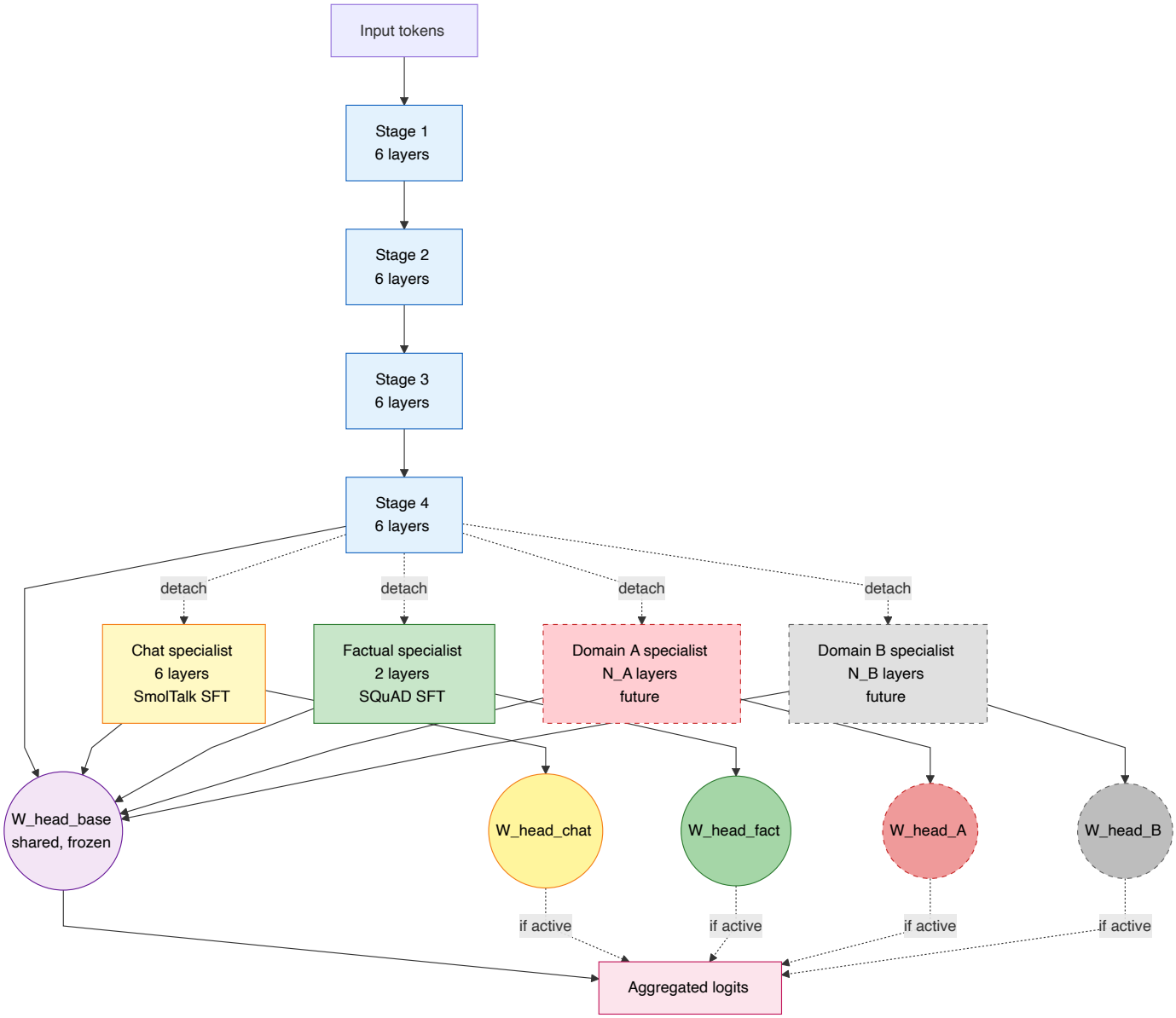
Figure 6.4 (Heterogeneous specialist ecosystem on shared base). Frozen base (S1–4, blue) as stable API. Chat specialist (yellow, 6L, SmolTalk) and factual specialist (green, 2L, SQuAD) attach with dedicated heads. Both project through shared $W_{\text{head,base}}$ + dedicated head; Stage 4 prefix bit-identical across both.

6.6.8 Implications beyond the specialist ecosystem

Two consequences: **(i)** base stages need not be uniform depth — factual specialist shows a functionally complete stage can be 2 layers; a base with heterogeneous stages (deeper first, progressively shallower) is concealed by uniformity-by-default. **(ii)** stage boundaries may be determined dynamically rather than fixed a priori. Both natural extensions for a companion paper (§10.3).

6.6.9 Paper-level contributions of this experiment

(1) Same-base different-SFT causal test: chat-SFT objective causes gullibility (4–6× reduction). (2) Specialist ecosystem empirically supported: two specialists on same base deployable per-task. (3) Stage 5 size not critical: 2 layers works where 6 failed at the same data scale. (4) Data-size threshold: 14K fails, 3M succeeds; ~1M tokens per specialist minimum. (5) Retrieval trust is a learnable scalar. (6) Resume feature production-validated.



6.6.10 What this experiment does not establish

Compute-matched comparison (chat 22M vs factual 12M tokens); whether 0-layer head-only specialist matches 2-layer behavior; generalization to other SFT objectives; multi-specialist ensemble behavior (follow-up in §10.2).

---

7. On-Device Deployment: Apple Silicon Verification

7.1 MLX Port and Benchmark Setup

We ported the 1.3B checkpoint to Apple's MLX framework for benchmarking on M1 Ultra (64 GPU cores, 128GB unified memory). The MLX port produces fp32 outputs bit-identical to PyTorch reference. PyTorch MPS baseline on the same hardware provides cross-validation. Autoregressive decoding on 9–41 token prompts; `mx.metal.synchronize()` before timing.

7.2 Architectural Speedups Verified on Consumer Silicon

Operation	MLX (ms)	PyTorch MPS (ms)	Speedup vs full d24
Full 4-stage (d24)	24.7	23.0	baseline
Stage 1 only (prefix pruning)	8.5	9.0	**~2.7× faster**
Stage 1→2 prefix	14.0	12.9	~1.8× faster
Stage 1→3 prefix	21.0	17.8	~1.3× faster
Skip S1→S[k] (skip-stage)	12.6	12.4	~1.9× faster
3-branch parallel (default dispatch)	36.9	35.1	**0.65× (slower)**

Stage 1 alone runs in 8.5–9.0 ms vs 24.7 ms full-stack — 2.7× faster. Architectural speedups aren't datacenter-hardware peculiarities.

7.3 Framework-Agnostic Consistency

Qualitative pattern identical across MLX and PyTorch MPS. Absolute latencies differ 5–15%; relative speedups match within noise. Benefits come from model structure, not specific runtime — would transfer to llama.cpp, Core ML, etc.

7.4 Parallel Stage Dispatch via MLX Streams

Default single-queue dispatch serializes parallel branches (§7.2 3-branch at 0.65×). MLX's per-stream (`mx.stream()`) assigns each branch its own command queue:

N branches (each full d24)	Default stream	Per-stream	Speedup from streams
2	1.97× single	1.51× single	**1.30×**
3	2.94× single	2.62× single	1.12×
4	3.88× single	3.24× single	1.20×

2-branch benefits most (each uses ~half of 64 cores). At 3–4 branches, GPU saturates. Full theoretical speedup ($T_{S_1} + \max(T_{S_k})$) requires multi-device dispatch.

7.5 Deployment Positioning: Datacenter BP, On-Device PoE

Task-polarized CORE results (§5.5) suggest a **division of deployment contexts**.

Datacenter inference remains well-matched to BP. Cloud-scale amortizes global gradient cost across batch throughput; BP's factual advantage is directly consequential. Not a BP replacement here.

On-device inference is structurally different:

Constraint	Datacenter	On-device
Memory	Hundreds of GB per node	4–32 GB shared with OS
Compute budget per query	Amortized across batch	Single-user, real-time
Power/thermal	Plug-in, liquid-cooled	Battery, passive cooling
Model update	Centralized, continuous	OTA, infrequent, bandwidth-limited
Device heterogeneity	Controlled	Phone SoCs vary by orders of magnitude
Factual breadth	Must cover everything	Retrieval-augmented from cloud/local

PoE's properties map directly: prefix pruning (1.3B → 325M at 87.5% factual), WAND (1.82×, 45% energy), speculative decoding (1.87× without separate drafter), post-hoc specialists (OTA add-ons without base retraining), retrieval as structural counterpart (§8.2.1's 712× gain). Device-local models have always relied on external retrieval for factual breadth — PoE's architectural requirement and on-device conventions coincide.

Paradigm. Datacenter BP = "one large brain"; on-device PoE = "reasoner with tools". Alignment of architecture to constraint, not compromise. Bridge: distilling BP teacher into PoE student for workloads where retrieval is infeasible (§10.2).

Positioning, not claim. Architecturally well-suited to on-device based on measured properties; full deployment claim requires device-level benchmarking across phones/tablets/laptops and comparison against quantization, distillation, small-from-scratch, MoE.

---

8. Discussion

8.1 The Quality-Gap Trajectory: Non-Compressing Structural Floor

Configuration	Model	Training ratio	Final BPB	Gap vs matched BP
Per-layer PoE, 30M, WikiText-2	GPT-2 (6L, 384d)	—	—	12%
Clustered PoE (5L/stage), 897M, ClimbMix	GPT (20L, 1280d)	12	0.786	6.6% (vs 0.737)
Clustered PoE (6L/stage), 1.3B, ClimbMix	GPT (24L, 1536d)	**10**	**0.743**	**6.0% (vs 0.701)**
Clustered PoE, 1.3B from-scratch at r=20	GPT (24L, 1536d)	**20**	**0.720935**	**6.52% (vs 0.676788)**

Three observations: per-layer → clustered halved gap (12% → 6.6%) via coarser factorization; 897M → 1.3B at comparable ratios did not materially compress (6.6% → 6.0%); r=10 → r=20 shows **no compression** (6.0% → 6.52%). The flat budget response rules out strong H-E (gap does not shrink toward zero with more training); the persistent gap across three budgets consistent at ~6% with slight widening rules out weak H-S (no floor at all). The evidence sits on a **bounded structural floor** near 6–6.5% for this architecture family.

Within-run trajectory reinforces H-S interpretation. The r=20 gap is not stable — it widens convexly from ~4.3% (warmup) through 5.83% (step 20K) to **6.52% (final)**, with 31% of the total +2.20pp widening concentrated in the final 6K warmdown steps. BP's advantage emerges precisely during fine-grained optimization where global gradient coordination differentiates from local gradients. This is the signature of a structural rather than budget-dependent gap: a genuine budget artifact would shrink as both models converge; a convex-widening pattern during the phase most sensitive to optimization precision confirms BP is exercising a capability PoE cannot access.

Structural-floor interpretation. Combining the non-compressing r=10 → r=20 response with the convex warmdown-widening pattern, the evidence supports **H-S (structural floor)**: a bounded architectural cost of local learning, not a training-budget artifact. Best characterization of the current data: "PoE pays ~6–6.5% quality floor for local learning." The r=30 from-scratch experiment (§10.1) directly discriminates whether a weak slow-compressing H-E component exists alongside H-S.

Scale-dependent knowledge localization. Independent of gap trajectory:

Scale	Stage 1 factual accuracy	Stage 4 factual accuracy	First-to-last gap
897M, ratio 12	62.5% (5/8)	87.5% (7/8)	25pp
1.3B, ratio 10	**87.5% (7/8)**	**87.5% (7/8)**	**0pp**

At 1.3B Stage 1 alone recalls the same facts as the full model. Deeper stages become factually redundant — contribution is statistical refinement without additional knowledge retrieval. WAND calibration corroborates: p99 deltas decrease monotonically (7.09 → 3.03 → 2.15).

A 1.3B PoE deploys as a **325M-parameter factual retriever** at 25% compute with 87.5% full-model accuracy — Stage 1 trained as a complete predictor from the start. WAND extends to 1.82× wall-clock at 100% top-1 agreement. **Architectural properties are independent of the quality gap** — they hold whether H-S, H-E, or H-O is the correct interpretation.

8.2 Connection to Biological Learning

Three correspondences link clustered PoE to biological cognition: **structural** (organization), **functional** (capability profile), **compensatory** (external augmentation). Convergent-evolution inference — locality as a constraint predicts the capability profile and compensatory pattern in both systems.

Structural. Neocortex is organized as locally-connected regions with dense internal / sparse inter-region communication — neither independent neurons nor globally-coordinated system. Clustered PoE occupies this middle ground.

Biological Structure	PoE Analog
Cortical column (~0.5mm)	Single transformer layer
Cortical area (V1, V2, ...)	Stage (cluster of layers)
Intra-area connectivity	Intra-stage gradient flow
Inter-area projections	Detached forward activations
Hebbian/STDP plasticity	Local cross-entropy loss

Topology common to both (dense local, sparse inter-region, no global gradient); neuromodulation, predictive coding, oscillatory binding not modeled.

Functional. Biological cognition is strong on pattern, generalization, gist; weak on verbatim. CORE shows PoE with the same profile: strong on PIQA (+5.0pp), CSQA (+5.8pp), BigBench CS Algorithms (+11.4pp); weak on Jeopardy (−81%), SQuAD (−44%), LAMBADA (−30%). Rare events provide weak local gradient under any local objective.

Compensatory. Biology does not close the verbatim gap by architecture change — it compensates externally (writing, reference, search). Extended Mind (Clark & Chalmers, 1998) and distributed cognition (Hutchins, 1995) formalize this. Direct implication: rare-fact deficit is a structural property to address through external retrieval. PoE + retrieval mirrors the biological configuration.

Correspondence	Biology	Clustered PoE	Shared constraint
Structural	Cortical regions, no global error	Stages + detach, no cross-stage gradient	Locality of learning signal
Functional	Strong pattern; weak verbatim	Strong commonsense; weak rare-fact	Local objectives learn local statistics
Compensatory	External tools	External retrieval (RAG)	Complete = internal + external

BP's monolithic approach is out of reach for any system without global gradient coordination (biological, on-device). PoE aligns with the distributed-cognition paradigm.

8.2.1 Empirical validation of the retrieval structural argument

§8.2 predicts rare-fact failures from PoE's locality profile should be recoverable through RAG without retraining. Tested on dual-head d30 step-5557 against §6.3's unreachable knowledge: "The first US president was" places `Washington` at rank 259 / 1377 / 1462 under S4-base / S5-dual / S5-base — effectively unreachable. Probes prepend passages (Wikipedia, bulleted, dialog, inline quote, short fact, prefix-only); controls (weather, math, adversarial wrong-fact under six authority framings); cross-topic sweep (Washington, Lincoln, Einstein, Beethoven, Curie); BP baseline included.

R1 — RAG recovers rare-fact retrieval under correct passages. Under S5 dual with bulleted passage, multi-token greedy produces `George Washington` with teacher-forced $P = 0.4272$ — **712× over no-RAG** (0.0006). Wikipedia 0.3608. Format ranking: bulleted ≈ Wikipedia > dialog ≈ inline-quote > short-fact > prefix-only (matching SmolTalk).

Single-token probes: `George` rank 1 (P = 0.428) but `Washington` rank 14 (P = 0.007). Methodological asymmetry — first-name is an early high-probability decision; surname requires conditioning on first-name output. Multi-token greedy chains through the first-name decision, turning a difficult surname probe into a reliable conditional lookup. **Sequence probability, not single-token surname, is the deployment-relevant quantity.**

R2 — Controls confirm content-based retrieval use.

Passage	Washington rank (S5 dual)	Top-1
Baseline (no passage)	1377	`<|assistant_end|>`
Irrelevant (weather)	1161	`ass`
Irrelevant (math)	985	`born`
Adversarial (`John Snow was the first US president`)	44	`John`
Correct (Wikipedia)	**14**	`George`

Irrelevant passages keep rank above 900 — recovery is content-driven, not length/positional. Adversarial wrong-fact shifts top-1 to `John` but Washington still rises to rank 44 — model weighs retrieval against learned priors.

R3 — Cross-topic generalization with baseline-prior-dependent recovery.

Query	Target	Baseline rank	+RAG rank	Improvement
"The 16th US president was"	`Abraham`	7238	134	54×
"The 16th US president was"	`Lincoln`	1164	14	83×
"The composer of the Ninth Symphony was"	`Ludwig`	105	2	52×
"The composer of the Ninth Symphony was"	`Beethoven`	57	3	19×
"The theory of general relativity was developed by"	`Albert`	1	1	already-known
"The theory of general relativity was developed by"	`Einstein`	2	2	already-known

Where baseline is unreachable (Lincoln, Ludwig, Beethoven), RAG recovers 19–83×; where already top-1 (Einstein), RAG provides no lift. Retrieval completes the rare-fact path locality-constrained training cannot encode densely.

R4 — RAG safety is retrieval-quality-bound, not architecture-bound. Cross-topic wrong-fact sweep. Metric: $P(\text{correct})/P(\text{wrong})$, higher = more robust.

Topic	BP baseline	Stage 4 prefix	S5 dual (std fwd)
Washington	0.12	0.09	0.02
Lincoln	0.26	0.27	0.06
Einstein	203	7,284	**40,990**
Beethoven	0.01	0.01	0.00
Curie	**1,455**	12	0.88
Cross-topic median	**0.26**	**0.27**	**0.06**

Three topics (Washington, Lincoln, Beethoven) catastrophically gullible under every strategy; two (Einstein, Curie) robust across every strategy. Architecture effect non-uniform: BP more robust on Washington/Curie; S5 dual on Einstein. Median: BP 0.26, S4 prefix 0.27, S5 dual 0.06 — SFT-amplified standard forward ~4× more gullible in median.

Primary determinant is **baseline prior strength**, not architecture:

Topic	Baseline P(correct) under S5 dual	Wrong-RAG ratio
Einstein	0.648 (rank 1)	40,990
Curie	0.006	0.88
Washington	0.001 (ambiguous prompt)	0.02
Lincoln	0.000	0.06
Beethoven	0.002	0.00

Strong prior (Einstein): wrong retrieval cannot shift the distribution. Weak prior: retrieval dominates regardless of direction. **RAG gullibility is a function of baseline prior strength, retrieval quality, and prompt disambiguation — architecture is a distant fourth.**

R5 — Textual skepticism markers do not calibrate the model. Six authority framings of the same wrong claim produce near-identical P(wrong-name):

Passage framing	P(George)	P(John)	Ratio
Assertive ("Fact: ...")	0.007	0.364	0.02
Matter-of-fact	0.016	0.471	0.03
Hedged ("Some sources say...")	0.022	0.430	0.05
Hedged + disputed	0.017	0.477	0.03
Explicitly flagged as wrong	0.012	0.356	0.03
Internal contradiction (both names)	**0.053**	0.485	**0.11**

Textual markers produce essentially no calibration benefit — P(wrong-name) stays in 0.36–0.48. "Warning: this claim is disputed" with the wrong assertion produces P(John) = 0.356, indistinguishable from confident assertion. Internal contradiction helps the correct name (5×) but wrong still dominates. **Model-side self-calibration is not a viable safety layer**; deployment requires external retrieval quality control.

R6 — Inference-time escape hatch as PoE-specific capability. Under RAG with Wikipedia passage:

Branch	`George` rank / P	`Washington` rank / P
S4 base (prefix pruning)	1 / 0.335	**11 / 0.015**
S5 dual (standard forward)	**1 / 0.362**	14 / 0.007
S5 base (drift-through-base)	1 / 0.008	45 / 0.003

S5 dual strongest under trusted retrieval ($P(\text{"George Washington"}) = 0.43$); S4 prefix close second (0.31). Under adversarial retrieval, wrong-RAG median (S4 prefix 0.27 vs S5 dual 0.06) identifies S4 prefix as safer fallback. Dual-head PoE's architectural value: *inference-time switching* without retraining. BP has no equivalent post-hoc switch.

Implications. Four composing recovery mechanisms — prompt reformulation, inference-path selection, retrieval augmentation, multi-token decoding. §7.5's deployment empirically validated under trusted retrieval. §10.2 item (5) elevated from prediction to preliminary result. §10.4 gains: **RAG safety requires external retrieval quality control**. PoE's inference-time escape hatch is a concrete architectural benefit over BP not captured by BPB.

Summary. §8.2's structural claim is empirically anchored under correct retrieval and scoped under adversarial retrieval. 712× sequence-probability gain, generalization across four rare-fact topics, control-validated content-based retrieval demonstrate PoE+retrieval is not aspirational. Scope boundaries identify deployment requirements, not failures of the structural claim. All measurements on step-5557 with Stage 4 delta = 0.0000.

8.3 Why PoE Succeeds Where Prior Local Learning Fails

Prior local learning (target propagation, greedy layerwise, HSIC, Forward-Forward, decoupled greedy) never achieved parity with BP at production scale. Common strategy: **introduce an indirect signal** (synthetic targets, kernel proxy, goodness scores) to substitute for BP's chain rule.

PoE differs at the starting point: **(1) Objective identity** — each stage optimizes the same CE as the global task; no orthogonal-to-global minimization. **(2) Shared projection as implicit coordinator** — weight-tied W_{head} receives gradients from every stage loss, enforcing common geometry without auxiliary network. Distinguish **compute coupling** (BP: each unit's execution depends on another's) from **parameter sharing** (PoE's shared head): PoE inherits all systems properties BP's compute coupling prevents. W_{head} must be frozen during post-hoc SFT. **(3) Direct label signal** — every stage

computes loss against ground truth; no information loss in transit.

These follow from Hinton's (2002) PoE: $\log p_{\text{PoE}}(\mathbf{y} \mid \mathbf{x}) = \sum_k \log f_k(\mathbf{y} \mid \mathbf{x}) - \log Z(\mathbf{x})$ mechanically yields same-loss per expert, shared normalization, direct likelihood. Methods asking "how to replace BP's pieces" accumulate indirections; methods asking "how to factor the task locally" arrive at mechanisms whose parts derive from a different decomposition.

8.4 Multi-Head Attention Parallel: A Structural Principle

The dual-head construction parallels the design that made attention viable. Vaswani et al. (2017): "multi-head attention allows the model to jointly attend to information from different representation subspaces; with a single attention head, averaging inhibits this." Theorem 2.4.1/Remark 2.4.2 strengthen analytically: attention is context-dependent PoE/Log-OP; multi-head is an ensemble of parallel PoE aggregators. Post-training analyses (Clark et al. 2019; Voita et al. 2019) confirm individual heads specialize.

$W_{\text{head,base}}$ is asked to project the final stage into a distribution that must simultaneously express rare-fact retrieval, commonsense, code, chat, creative — qualitatively distinct geometries. Single-projection inadequacy is structurally identical at the output and at attention.

Three retrospective predictions. (1) §6.3 head-destruction — single head fine-tuned on a new distribution shifts toward the new and away from the old (Washington −14.7), exactly what multi-head motivation predicts. (2) Non-emergence of stage-specific head partitioning — all stages gradient-contribute to the same matrix, requiring *architectural* separation. (3) §6.4 capacity trade-offs — SFT gain on code/creative at cost of math is what single-projection bottleneck produces.

Same structural solution. Multi-head: H separate $(W_Q^{(h)}, W_K^{(h)}, W_V^{(h)})$. Dual-head: dedicated $W_{\text{head,k}}$ per specialist + frozen $W_{\text{head,base}}$ combined additively. Principle identical: *multiple projections, each in its own subspace, combined for a richer joint distribution*. Standard Transformers multi-head within attention but collapse to single projection at the output. Pretraining on monolithic loss doesn't surface the need; post-hoc specialization does. §6.5.5's +2.4 logit gain is Theorem 2.4.4 at the output layer.

8.5 Infrastructure Fragility and the Locality Imperative

Frontier training emphasizes engineering — GPU memory resembling pre-virtual-memory era; all-reduce scaling with global batch; dataset replication per node; GPU failures at significant rates. BP's global sync makes each a first-order bottleneck: one straggler slows the step, one failed link stalls the collective, one OOM terminates the job. Clustered PoE suggests an alternative: **reduce dependence on failure-free global coordination** — slow stage slows only itself; failed device affects only hosted stages; memory fragmentation bounded.

End-to-end fault tolerance under adversarial injection is future work, but the precondition (layer independence) is established; 1.3B ran on non-RDMA networking (AWS ENA) without instability. A single observed crash — disk-full recovered from checkpoint with one iteration lost — illustrates PoE-compatible failure character: mundane exhaustion, not synchronization cascade. Locality may be the honest response to physical limits on global coordination that frontier training is already encountering.

High-latency network tolerance — a precursor signal for cross-AZ feasibility. The 1.3B ENA run operated at ~10–25 Gbps per-node bandwidth with higher latency than EFA's single-AZ RDMA (400 Gbps); throughput degraded ~3× (MFU 13% vs 40% expected with EFA) but convergence was unaffected and training completed without synchronization instability. ENA is the network class AWS provides for cross-AZ communication, and the observation that local learning tolerates ENA-class conditions is a precursor signal for cross-AZ feasibility — a regime where BP's global all-reduce faces severe step-time penalties that scale with the slowest cross-zone link. Industry GPU supply is increasingly multi-AZ (H100/B200 capacity distributed across availability zones due to datacenter space, power, and network constraints); locality-tolerant training is structurally positioned to unlock consolidated use of geographically distributed accelerators that BP treats as incompatible with single-cluster training. Current evidence is a network-class tolerance demonstration rather than a true cross-AZ benchmark — formal cross-AZ comparison against BP baseline is future work (§10.5 item 5).

8.6 Refuted Mechanistic Hypotheses

Two hypotheses tested and refuted. **Emergent head partitioning** (that W_{head} self-partitions into stage-specific row-subspaces) falsified by §6.2 — Stage 5 head updates propagated to all frozen stages (max logit diffs 17–20, top-1 flips on 2/3 probes). Partitioning cannot emerge through gradient dynamics alone because all stages contribute to the same matrix. **Projection-robustness** (that PoE training through a moving head produces perturbation-robust representations) falsified by Gaussian-perturbation sweep: at every $\sigma \in \{0.01, ..., 1.0\}$, PoE degrades as much as or more than BP (ratio 1.02–2.59×). The correct mechanism behind PoE's architectural properties is **per-stage detachment itself**.

---

9. Limitations

1. **Two clustered configurations at production scale**: `poe_every=5` on d20 (897M, r=12), `poe_every=6` on d24 (1.3B, r=10 and r=20 from-scratch). Other stage sizes, depths, ratios unexplored.
2. **Quality gap does not compress across training budgets**: 1.3B r=10 = 6.0%, r=20 final = 6.52% (no evidence of budget-driven convergence). Within the r=20 run, gap widens convexly through warmdown (+2.20pp over 26K steps, with 31% concentrated in the final 6K warmdown steps), consistent with structural component. Evidence supports H-S; r=30 remains the test for any slow-compressing H-E component.
3. **CORE gap is task-polarized** (§5.5): rare-fact retrieval vs commonsense / algorithmic patterns. Single-composite reporting misrepresents the method.
4. **PoE-specific weaknesses at 1.3B base**: code collapse, associative interference, short-prompt degeneracy. §6.4 revised: degeneracy not PoE-specific (56% PoE, 42% BP); code and associative interference resolved by SFT — training-regime artifacts.
5. **Rare-fact retrieval is a PoE weakness**, not an advantage. Stage 1's 87.5% factual accuracy is structural; does not imply CORE-scale superiority.
6. **No PoE-specific hyperparameter tuning**: ratio, LR, loss weighting, warmdown all inherited from BP baseline.
7. **Warmdown schedule** designed for BP; may not be optimal for per-stage dynamics.
8. **Continual learning hypothesis untested at production scale**; only Split-CIFAR-10 (Appendix A).
9. **Infrastructure robustness preliminary** — §8.5 fragility discussion is design-motivated; fault-injection future work.
10. **No direct token-level MoE comparison** at matched compute.
11. **Training-free ensemble heuristics do not outperform sum** (Appendix C.5). Distinct from §6.5.5's parallel-composition gain, which is an architectural property of dual-head — no training-free reweighting recovers what dual-head SFT introduces.
12. **Dual-head validation scope.** 5557-step compute-matched validation confirms Stage 4 preservation (delta = 0.0000 across twelve checkpoints). Predictions (b) capacity trade-off mitigation and (c) chat/code parity untested. Generation: 0/11 confabulations (dual) vs 2/11 (v2).
13. **Modular update discipline untested** — §10.4 is a design proposal.
14. **Parallel-composition gain measured on a single fact** (Washington, +2.4 logit). Mechanistic prediction says generalize; single-point measurement.
15. **Multi-specialist composition empirically untested**: §6.6 shows $S = 2$; multi-branch at $S > 1$ (Figure 10.1), interference, routing policies open.
16. **Production-scale limited to Transformer LM** (modified nanochat). Cross-architecture validated only at small scale.
17. **Systematic BP-hyperparameter asymmetry** (§10.1 H-O). All settings tuned for BP dynamics. **The measured quality cost is an upper bound on PoE's intrinsic cost.** Thesis-relevant results (modularity, architectural properties, dual-head preservation) are independent of training-hyperparameter tuning.
18. **RAG empirical scope.** §8.2.1: (a) single checkpoint d30 step-5557; (b) single-source only; (c) no BP+SFT comparison; (d) format sensitivity may be SmolTalk-specific; (e) CORE-scale RAG validation pending.

---

10. Future Work

10.1 Distinguishing Structural, Engineering, and Optimization Causes

Three non-exclusive hypotheses for the quality-cost gap: **H-S (Structural)** — per-stage supervision intrinsically cannot recover information BP propagates; no training/tuning closes the gap below some floor. **H-E (Engineering)** — per-stage signal is weaker; PoE's compute-optimal ratio exceeds BP's; additional training closes most of the gap. **H-O (Optimization)** — gap reflects systematic BP-hyperparameter asymmetry (§9 item 17); PoE-specific tuning at fixed budget could close it.

Evidence supports H-S with no detectable H-E component at current budgets. Two empirical observations from final data: (i) r=10 → r=20 shows no compression (6.0% → 6.52%), inconsistent with any strong H-E prediction; (ii) within the r=20 run, gap widens convexly through warmdown (+2.20pp over 26K steps, with 31% concentrated in the final 6K warmdown steps, 4.32% → 6.52%), tracking BP's exercise of global-coordination capability in fine-grained optimization. Both patterns point toward a bounded structural floor. The strong convergence interpretation (H-E dominant) is inconsistent with the data. Whether a weak slow-compressing H-E component exists on top of the H-S floor is still open — r=30 is the test.

Two direct tests: (1) **r=30 from-scratch training** (decisive test for slow H-E component) — if gap stays at ~6.5%, H-S dominates exclusively; if gap drops meaningfully (e.g., ≤5%), a slow-compressing H-E component is present alongside H-S. (2) **Unified per-stage heads at r=20** (§10.2 item 6, H-O) — tests whether architectural refinement at the output layer closes part of the floor.

Secondary follow-ups: stage-specific fine-tuning (freeze S1–3, continue only S4); `poe_every ∈ {3, 4, 8, 12}` ablation; PoE-adapted LR schedule and warmdown; optimizer ablation (Muon vs AdamW vs Lion vs Shampoo); 3B/7B scale at r=20 to test whether the floor compresses with parameter count.

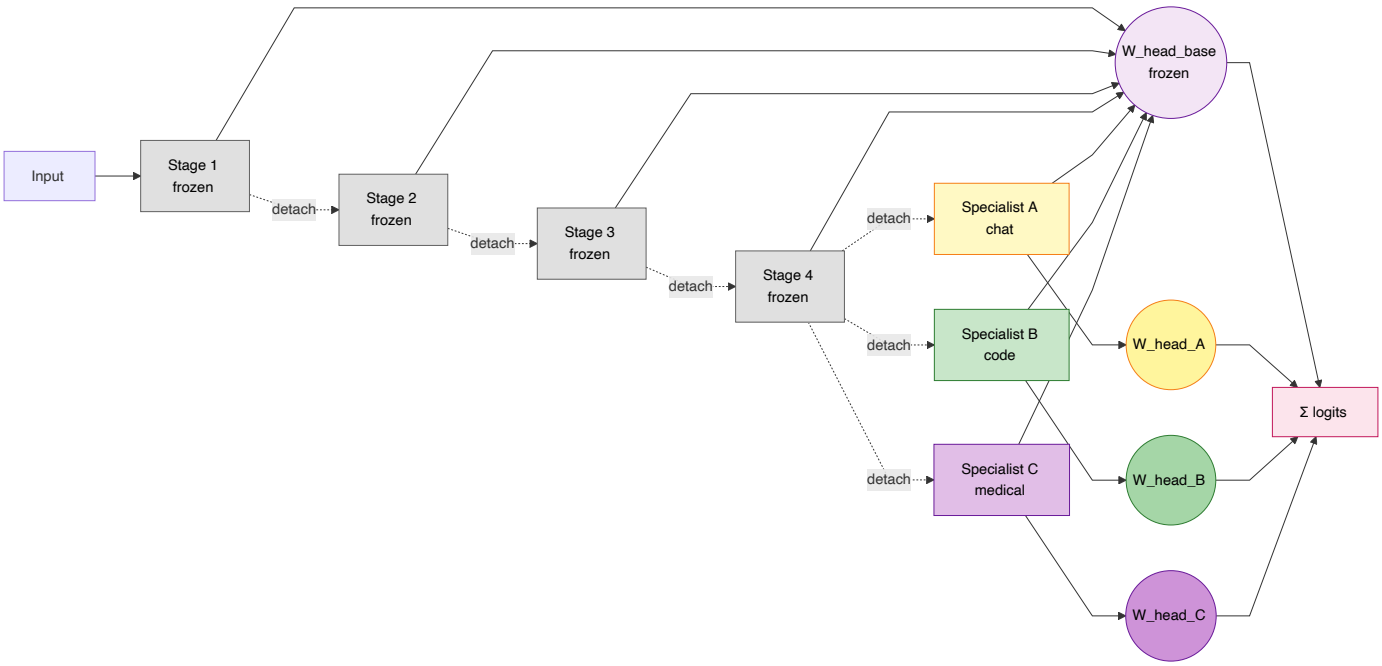
10.2 Exploiting PoE's Architectural Advantages

Six concrete follow-up experiments prioritized by expected information value.

(1) Multi-specialist extension of dual-head. Natural extension from single specialist (§6.5) to multiple simultaneously-active specialists, each trained independently with its own head:

$$\text{logits}(x) = \sum_{k=1}^{4} h_k W_{\text{head,base}} + \sum_{s \in \text{active}} h_s \left(W_{\text{head,base}} + W_{\text{head,s}} \right) \tag{12}$$

Figure 10.1 (Stage-level MoE with dual specialist heads). Frozen base stages (grey) and frozen base head (purple) = shared backbone. Specialist stages (yellow/green/magenta = chat/code/medical) trained independently on domain-specific data with dedicated heads. Router selects active subset at inference; inactive specialists omitted from summation.



> Invariants: (i) per-stage detachment across specialists (no inter-specialist or specialist→base gradient); (ii) base head participates in every route, preserving factual projection directions; (iii) each $W_{\text{head,s}}$ in an independent subspace — direct parallel to multi-head attention (§8.4); (iv) activation is post-training — no retraining to change composition.

Stage-level MoE with per-specialist projection geometry. Open: can independent specialists train without interference? does multi-branch parallel composition show additive gains?

(2) BP teacher → PoE student distillation. Distill BP teacher via per-stage CE against soft targets. Measures whether teacher's rare-fact sharpness transfers.

(3) Learned stage routing. Small MLP router selecting which stage to trust, using validation prompts with ground-truth paths. Training-free routing failed (Appendix C.5); learned routing is the apparent path.

(4) Stage-level MoE empirical validation. Systematically train domain-specific specialists, measure multi-active composition quality, compare against equivalent-parameter token-level MoE.

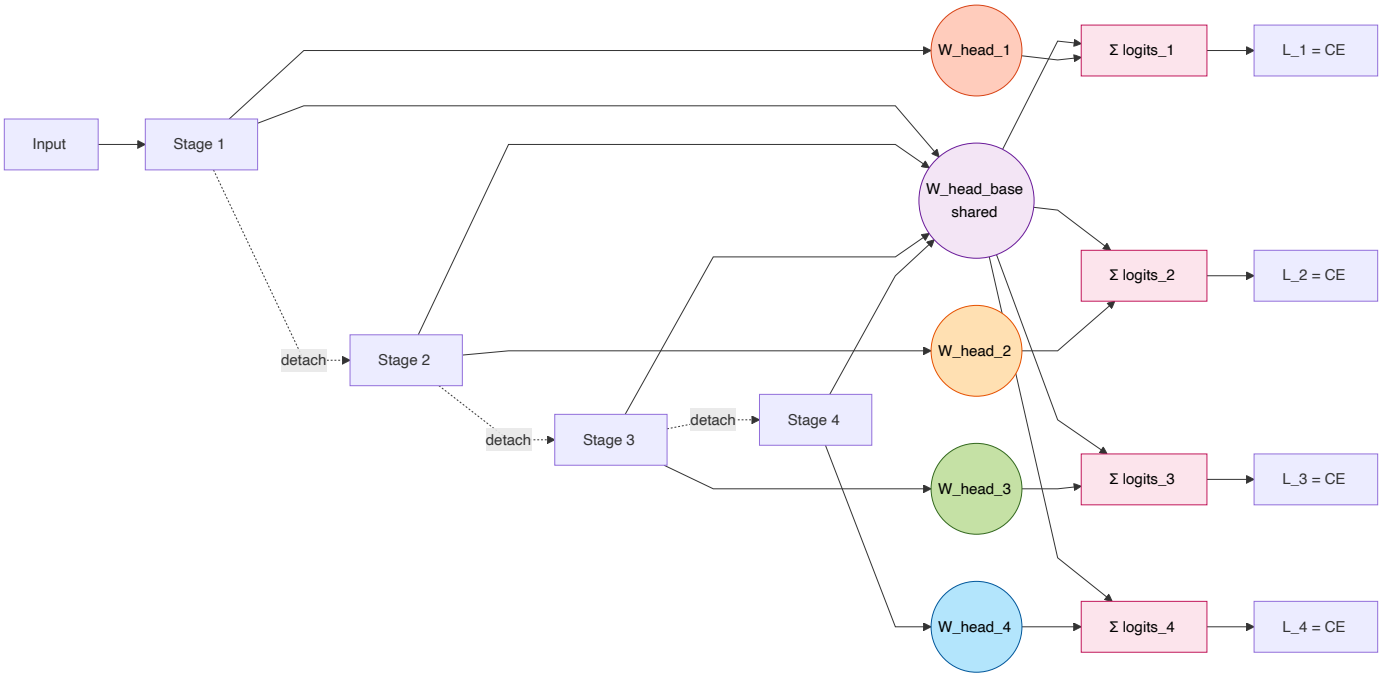
(5) RAG-augmented PoE — CORE-scale validation. §8.2.1's four-part validation on step-5557. Follow-up: CORE rare-fact subscores under RAG; multi-source retrieval with contradiction detection; retrieval-format sensitivity across SFT datasets; BP+SFT comparison to isolate whether gullibility is PoE-SFT-specific.

(6) Unified per-stage head architecture. Apply the "shared + dedicated head" template uniformly — each stage k receives its own $W_{\text{head,k}}$:

$$\text{logits}_k = h_k(W_{\text{head,base}} + W_{\text{head,k}}), \quad \mathcal{L}_k = \text{CE}(\text{logits}_k, y), \quad \text{logits}_{\text{full}} = \sum_k h_k(W_{\text{head,base}} + W_{\text{head,k}}). \quad (13)$$

Architectural realization of the multi-head principle (§8.4) at the output layer. Dual-head (§6.5) sparse, multi-specialist partial, unified full — points on a sparsity spectrum.

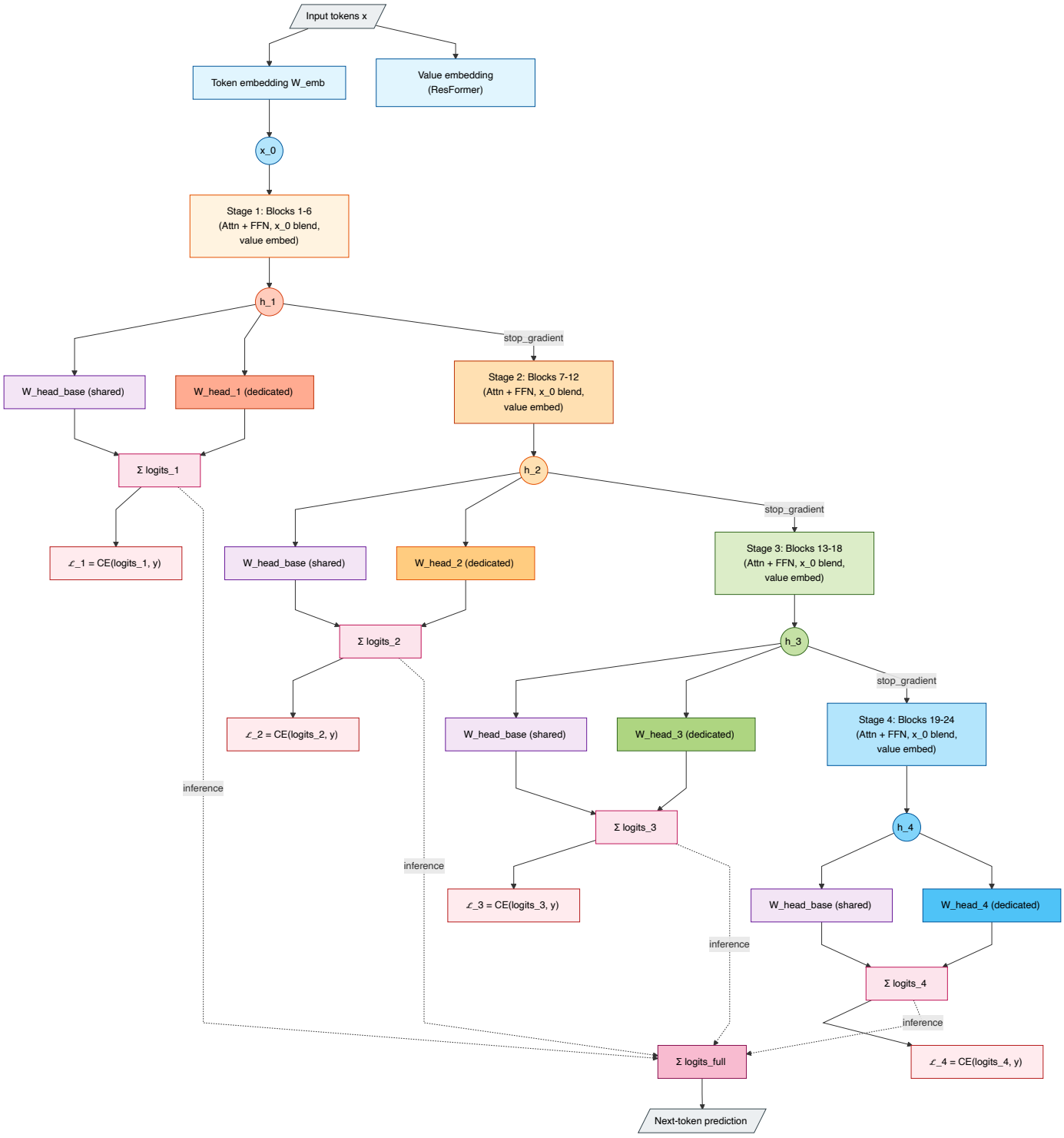
Figure 10.2 (Unified per-stage head architecture; proposed). Each stage projects through shared $W_{\text{head,base}}$ (purple) and dedicated $W_{\text{head,k}}$ (colored). Boundaries detached.



> Invariants: (i) per-stage detachment preserved; (ii) $W_{\text{head,base}}$ receives gradient from every stage's loss as a universal LM prior; (iii) each $W_{\text{head,k}}$ updated only by its own stage; (iv) parallel to multi-head attention — shared + per-stage at output ≡ per-head (Q, K, V) with shared W_O.

Predictions: **(a) H-O evidence** if unified heads close a fraction of the 6.0% gap at r=10. **(b) Interpretable per-stage geometry**. **(c) Modular post-hoc adaptation** via per-head updates. Parameter cost for 1.3B ($d = 1536, V = 65{,}536$): each head adds ~100M (~30% across $K = 4$). Mitigations: tie to embedding; low-rank $W_{\text{head,k}} = A_k B_k$ ($r = 64$) → ~150× reduction; or restrict to deeper stages.

Figure 10.3 (Detailed per-stage architecture; proposed). Unified PoE for 1.3B ($K = 4$, $S = 6$, $d = 1536$). Transformer blocks per stage with x_0 blending; stop-gradient at boundaries; each h_k through $W_{\text{head,base}}$ (tied) + $W_{\text{head,k}}$.



> Four stages × six Transformer blocks. x_0 blending + ResFormer value embeddings reach every block. Boundaries apply `stop_gradient`. $\text{logits}_{\text{full}} = \sum_k \text{logits}_k$. Original clustered PoE: $W_{\text{head,k}} = 0$; dual-head: only $W_{\text{head},5} \neq 0$; unified: all four non-zero.

Lower-priority: per-category WAND calibration; tree-based speculative decoding (Medusa-style) beyond 1.87×.

10.3 Heterogeneous and Dynamic Stage Architectures (Companion Paper)

§6.6 shows specialist depth is task-dependent. Two extensions to the base warrant companion treatment. **Heterogeneous base stage sizes** — uniform $S = 6$ is a default; a deeper first stage (9 layers) for factual capacity followed by progressively shallower refinements (6, 4, 2) combined with unified per-stage heads is a natural alternative. **Dynamic stage boundary determination** — adaptive placement via per-layer CE contribution, or learnable soft-assignments under an information-

theoretic criterion. Full methodological development is the subject of a planned companion paper.

10.4 Modular Update Discipline

PoE's modular structure suggests a version discipline analogous to software framework versioning. **Minor versions** (continued pretraining, fixed hyperparameters, vocabulary, stage boundaries) preserve representation structure — compatible, like Python 3.11→3.12. **Major versions** (architectural/vocabulary changes, retraining from init) break compatibility — specialists migrate or retrain. Differs from the current paradigm where every release is effectively a major bump.

Compatibility adapters via head alignment. Major-version migration need not require full retraining. A specialist trained against base v1 projects through $W_{\text{head,v1}}$; vocabulary space is the shared frame. An adapter $A : \mathbb{R}^{d_{v1}} \to \mathbb{R}^{d_{v2}}$ minimizing $\|W_{\text{head,v1}}h - W_{\text{head,v2}}A(h)\|$ aligns spaces. Empirical support: Moschella et al. (2023) — relative representations invariant across models; Huh et al. (2024) Platonic Representation — correspondence increases with scale; Task Arithmetic (Ilharco 2023), Model Soups (Wortsman 2022) — linear composability. Bridging v1→v2 is coordinate alignment, not content relearning.

RAG deployment discipline. Three operational rules from §8.2.1: (i) retrieval source quality is the primary safety lever — textual markers produce no calibration benefit; (ii) inference-path routing is PoE-specific — dual-head supports per-query switching BP lacks; (iii) weak-prior topics require multi-source arbitration.

Proposal; adapter feasibility requires direct testing across base-version pairs.

10.5 Broader Exploration

(1) **Soft detachment** — learn per-boundary $\alpha \in [0, 1]$ interpolating full BP ($\alpha = 0$) and clustered PoE ($\alpha = 1$). (2) **Dynamic stage boundaries** via mutual information / CKA. (3) **PoE-specific optimizer** — per-stage LR, momentum, or asynchronous updates. (4) **Infrastructure fault-injection** testing §8.5 fragility claims. (5) **Interconnect sensitivity and cross-AZ training** — EFA vs ENA at matched scale; explicit cross-AZ deployment with BP baseline under matched cross-zone conditions to quantify locality's infrastructural advantage. (6) **Continual learning at scale** beyond Split-CIFAR-10. (7) **On-device runtime integration** (llama.cpp / Core ML) with energy and memory benchmarking. (8) **Mechanistic investigation** of residual failure modes.

---

11. Conclusion

PoE provides a unified foundation for local learning from small MLPs to production-scale LLMs.

Quality cost. 6.6% BPB at 897M (r=12); 6.0% at 1.3B (r=10); **6.52% at 1.3B (r=20 from-scratch, final)** — no compression from r=10 to r=20. Within the r=20 run, gap widens convexly through warmdown (4.32% at step 1k → 6.52% at step 26,430), with 31% of the +2.20pp total widening concentrated in the final 6K warmdown steps — BP's global coordination exercising its largest advantage in fine-grained optimization. CORE is **task-polarized** — PoE underperforms on rare-fact (Jeopardy −81%, SQuAD −44%, LAMBADA −30%) but exceeds on commonsense (PIQA +5.0pp, CSQA +5.8pp) and algorithmic (BigBench CS Algorithms +11.4pp).

Architectural advantages (unique to PoE). Stage prefix pruning at 25% compute with 87.5% full-model factual accuracy; WAND (1.82× wall-clock, 100% agreement); speculative decoding with Stage 1 as drafter (1.87×, 88% acceptance); post-hoc specialists via elastic depth. **Independent of the quality gap** — these hold whether H-S, H-E, or H-O is the correct interpretation.

Mechanism. Per-stage detachment itself drives the properties — each stage independent predictor via its own CE through a shared (not partitioned) head. Emergent head partitioning and projection-robustness refuted (§8.6). Multi-head attention parallel (§8.4) motivates dual-head. Compute-matched validation: Stage 4 rank-1 Washington bit-identical across 5557 steps; v2 destroys it by −14.73. Two-branch composition +2.4 logit over strongest single branch — Theorem 2.4.4 at the output layer.

Primary thesis: PoE is a principled architectural trade-off. Current evidence tentatively supports H-S (structural floor interpretation): the gap is a bounded cost of local learning rather than a vanishing budget artifact. Not a failed approximation of BP — a point in the design space with measurable cost and measurable advantages. Two follow-up experiments will sharpen the boundary: r=30 from-scratch (decisive H-S test) and unified per-stage heads at r=20 (H-O test).

Modularity validated. Six pillars: training-time modularity, inference-time composability, post-hoc extension, base preservation under SFT (delta = 0.0000), composable quality gains (+2.4 logit), heterogeneous specialist ecosystem (4–6× gullibility reduction). Modularity properties are empirically robust regardless of the structural-floor magnitude.

Deployment. On-device is PoE's natural habitat: memory, battery, heterogeneity align with prefix pruning, WAND, speculative decoding, elastic depth, modular update. Apple Silicon: 2.7× prefix speedup; 1.20–1.30× MLX streams. Datacenter deployment faces a known, bounded ~6.5% trade-off — not an eventual convergence to parity.

Biological correspondence. Triple correspondence — structural, functional, compensatory — evidences that capability profile follows from locality; "internal reasoner + external retrieval" is structural, not engineering convenience.

Research program. Production-scale neural-training component of a program unifying probabilistic foundations of retrieval, neural computation, and learning. Retrieval: Bayesian BM25 (Jeong 2026a), Vector Scores as Likelihood Ratios (Jeong 2026c). Neural computation: *From Bayesian Inference to Neural Computation* (Jeong 2026b) — attention is Log-OP/PoE, multi-head is parallel PoE ensemble, WAND is exact neural pruning; *Answer Bandwidth* (Jeong 2026d) — effective rank dominates activation performance. Shared: **structures become implicit statistical models; Bayesian inference composes their outputs into unified posteriors**.

The quality cost is bounded, not vanishing. The architectural advantages are unconditional. Whether the floor is exactly structural or has a slow-compressing component is a single r=30 experiment away.

Appendix

Appendix A: Cross-Architecture Small-Scale Experiments

Cross-architecture small-scale experiments (MLPs, CNNs, ResNets, Transformers on MNIST/CIFAR/WikiText-2) establishing scope-of-applicability outside Transformer LM. The framework applies wherever a model can be factored into independently-trainable sub-networks combined through a shared output projection.

A.1 Why Mechanical BP Replacement Fails

Three mechanical substitutions for BP — each removing or replacing a specific piece — tested on MNIST (784-256-128-10 MLP) and Tiny-ImageNet (VGG CNN): **Log-Odds Belief Propagation** (no sigmoid derivative): 10.28% (catastrophic); errors amplify 37× at layer 2, 846× at layer 1. **Local Autoencoders** (MSE reconstruction per layer): MNIST 90.99% (−6.93%), Tiny-ImageNet 0.83%. **Difference Target Propagation** (learned inverse networks): MNIST 97.70%, Tiny-ImageNet 14.02% (−15.18%). Local learning requires derivation from theory, not ad-hoc BP modification — the methodological observation motivating the PoE framing of §2 and §8.3.

A.2 MNIST, CIFAR-10, Transformers — Summary

Architecture	Task	Best PoE	Gap vs BP	Method
MLP	MNIST	98.00%	**+0.20%**	PoE + Diversity
VGG CNN	CIFAR-10	89.78%	−1.25%	Hierarchical PoE
ResNet-20	CIFAR-10	89.61%	−2.45%	Hierarchical PoE
GPT-2 (30M)	WikiText-2	60.66 PPL	+12%	Flat/Hier PoE

Flat PoE fails on hierarchical CNN features because the independence assumption is misspecified. Hierarchical PoE restores the conditional forward path while keeping independent per-layer losses. On Transformers, flat ≈ hierarchical because weight-tied embeddings create implicit gradient coupling — all expert losses contribute to the shared projection, providing inter-layer coordination that survives stop-gradient. Three alternative coordination methods (Difference Target Propagation, BYOL, Analytical Ridge) all converge to ~60 PPL, confirming the shared $W_{exthead}$ as the implicit coordinator (§8.3). Prefix Consistency (Prop 4.1) verified at three WikiText-2 scales with max PPL delta = 0.0000.

A.3 Continual Learning: Split-CIFAR-10

Method	Average Accuracy	vs EWC
Finetune	19.4%	1.0×
EWC	19.5%	1.0×
PoE + KD	**41.2%**	**2.1×**
PoE + KD + Retroactive Shrinkage	**45.5%**	**2.3×**

PoE+KD achieves 2.1× EWC's performance; retroactive shrinkage (inference-only, no retraining) adds +4.3pp, validating the Training-Inference Separation Principle (§2.3). Mechanism: in BP, new-task gradients flow through all layers, overwriting; in PoE, each layer's parameters update primarily from its own local loss, shielding earlier representations. Untested at production scale (§9 item 8).

Appendix B: 897M Scale Results

897M experiments preceding the 1.3B validation. Training on ClimbMix-400B at r=12 (5.22B tokens, 4980 steps, 8× A100 40GB NVLink).

B.1 BPB. Baseline 0.737, Flat PoE 0.786 (+6.6% gap). Dynamics: rapid increase to ~5% in steps 0–1000, then slow drift to 6.6%. Both improve at similar rates (~0.017 per 250 steps late), suggesting the gap reflects slower initial convergence rather than capacity difference. The 1.3B run (§5.1) reproduced a similar gap (6.0% at r=10); §5.6 showed the r=20 gap is **6.52% (final)** — confirming no budget-driven compression across three settings.

B.2 Training throughput. dt/step 2290 ms (BP) vs 3035 ms (PoE) — 1.33× overhead, below the theoretical 1.42× FLOPs prediction, likely from partial overlap of per-stage `lm_head` projections. **Pipeline-aware net throughput** for d48 with 4 stages: at 16 micro-batches, PoE net ratio 0.93×; at 32 micro-batches, 0.99×. For models requiring 4+ pipeline stages, clustered PoE achieves equal or better throughput.

B.3 Stage prefix pruning at 897M. Stage 1 achieves 62.5% (5/8); full 4-stage achieves 87.5% (7/8). At 1.3B, first-to-last gap collapses to zero — knowledge concentrates in Stage 1 as scale increases.

B.4 Parallel stage branching. Skip-stage S1→S3 (skipping S2) sometimes produces correct answers where full sequential fails. 3-branch ensemble (logit sum) blends properties of multiple paths. Empirical basis for stage-level MoE (§4.3, §6).

B.5 Stage repetition negative result. Duplicating Stage 1 at inference time produced small BPB degradation and no qualitative change. Stages are not composable through simple repetition — each stage's training-time input distribution is specific to the pretraining forward path.

B.6 Retracted 26-prompt qualitative claim. A 26-prompt evaluation suggested PoE ≥ BP on factual recall (4 PoE wins vs 3 BP wins). §5.5 CORE at n=500 refutes this: PoE underperforms BP on every factual-retrieval benchmark. The architectural distributed-storage observation (§5.2) survives; the quality-superiority claim from small samples does not.

Appendix C: Extended 1.3B Experiments

C.1 Parallel stage branching replication. The 897M skip-stage pattern (Appendix B.4) replicates at 1.3B: different branching patterns retrieve different facts. 3-branch parallel ensemble (logit sum) produces outputs blending three paths. Confirms stage branching as a real architectural primitive at 1.3B — basis for stage-level MoE.

C.2 Qualitative 23-prompt evaluation. BP 5 wins, PoE 2 wins, 16 ties. The 897M pattern does not reproduce at 1.3B; §5.5 CORE (n=500) is the primary evidence.

C.3 Systematic weaknesses at 1.3B base (pre-SFT). Three modes: (1) code generation collapse (0/10 at 1.3B vs 4/10 BP); (2) associative interference (Berlin Wall → 1889; first US president → Jefferson not Washington); (3) short-prompt repetition. §6.4 revised (1) and (3): code fully resolved by SFT (10/10); short-prompt repetition is not PoE-specific (56% PoE vs 42% BP base). Associative interference resolved at peak SFT (step 1600, Berlin Wall → 1989). Two of three are training-regime artifacts, not architectural ceilings.

C.4 Path-dependent factual recall. Different stage compositions retrieve different facts (see B.4 and C.1). Stage 1 alone retrieves some facts the full path cannot, and vice versa. Consistent with distributed knowledge storage hypothesis.

C.5 Confidence-correctness inversion. Training-free ensemble heuristics (max-probability, entropy-weighted, attention-based weighting) do not reliably exploit per-stage confidence; learned routing is the apparent path (§10.2 item 3). Distinct from §6.5.5's parallel-composition gain, which depends on dual-head-SFT-induced orthogonal posteriors.

Appendix D: SFT Training Procedure and Trajectory Detail

D.1 Setup. 6 new transformer blocks (layers 24–29) appended to the frozen 4-stage 1.3B base. ~325M trainable (~25% of base). SmolTalk dataset, 364M tokens, 1 epoch, best-fit packing. g5.12xlarge (4× A10G 24GB), 5557 steps, ~5h. LR 0.2× pretrain (matrix 0.004, unembedding 0.0016), warmup 5%, warmdown 50%.

Chat token warm initialization. Eight special tokens (`<|user_start|>`, `<|assistant_end|>`, `<|python_start|>`, etc.) warm-initialized from semantic counterparts (e.g., `<|user_start|>` ← "User").

Architecture depth-parity fix. Two features depend on total `n_layer`: `has_ve(i, n_layer)` and `backout_layer = n_layer // 2`. Fixes: absolute layer indices for `has_ve`; fix `backout_layer` to absolute or disable (trained `backout_lambda=0.0` confirms unused). Architecture-level, not PoE-specific — same issue under BP.

D.2 Warm init v1 failure → v2 fix. Without warm init, chat special token embeddings at norm ~30 (vs ~200+ for pretrained) are processed into meaningless representations by frozen Stages 1–4 — Stage 5 sees content but not structural markers. Result: fluent text, zero instruction-following ("What is DNA?" → health-supplement ad). Warm init from semantic counterparts resolves this by step 1000. Methodological guideline: elastic-depth SFT with new tokens requires warm init or pretraining inclusion.

D.3 Training trajectory across 28 checkpoints. Different capabilities converge at different rates and oscillate independently:

Capability	First emergence	Peak	Final (5557)
Chat register	Step 3200	5557	✅ Converged
Code generation (sqrt-optimized `is_prime`)	1600	2400	✅ Stable
Berlin Wall 1989	1600	1600	❌ Degraded to 1889
DNA explanation	1000	1000	❌ Oscillated
Arithmetic (2+2=4)	3400	3400	❌ Transient
Short-prompt coherence	3200	5557	🟡 Partial
"first US president" name	Never	—	❌ Unreachable (§6.3)

Observations: (1) no checkpoint optimal for all tasks — argues for task-specific Stage 5 specialists; (2) chat register and factual precision compete for Stage 5's ~325M budget — dual-head (§6.5) alleviates this; (3) capabilities shift rather than uniformly degrade; (4) some facts unreachable from Stage 5 — §6.3 head-destruction finding.

D.4 Shared head modification measurement (v1 with unfrozen $W_{exthead}$).

Prompt	Pre-SFT top-1	Post-SFT top-1	Max logit diff	Cosine sim
"capital of France"	Paris (10.9)	the (8.0)	18.9	0.88
"formula of water"	H (9.9)	H (5.8)	16.8	0.98
"What is DNA?"	DNA (10.8)	((7.8)	19.7	0.85

Top-1 flips on 2/3 prompts; logit diffs 17–20. Frozen Stage 1–4 representations are unchanged, but through modified head produce substantially different logits. Consequences: (1) $W_{exthead}$ must be frozen during post-hoc stage addition (§6.1); (2) frozen representations do not collapse under head modification — outputs remain fluent with top-1 changes, a general property of well-trained transformers.

D.5 1.3B weakness resolution summary.

Weakness	SFT resolution	Best checkpoint	Interpretation
Code collapse	✅ Resolved	1600+ (stable)	Training-regime artifact
Associative interference (Berlin Wall)	✅ Peak-resolved	1600 (1989 correct)	Capacity-dependent
Short-prompt repetition	🟡 Transformed	Character shifts	Partially addressable
(New) Head-destroyed factual knowledge	❌ Unresolved in v2	—	Caused by unfrozen head (§6.3); head-freeze predicted to resolve (§6.5)

References

- Belilovsky, E., Eickenberg, M., & Oyallon, E. (2019). Greedy Layerwise Learning Can Scale To ImageNet. *ICML 2019*.
- Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy Layer-Wise Training of Deep Networks. *NeurIPS 2006*.
- Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1), 7–19.
- Clark, K., Khandelwal, U., Levy, O., & Manning, C.D. (2019). What Does BERT Look At? An Analysis of BERT's Attention. *BlackboxNLP 2019*.
- Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, Ł. (2019). Universal Transformers. *ICLR 2019*.
- Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *JMLR* 23(120):1-39.
- Hinton, G.E. (2002). Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8), 1771-1800.
- Hinton, G., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7), 1527-1554.
- Hinton, G. (2022). The Forward-Forward Algorithm: Some Preliminary Investigations. *arXiv:2212.13345*.
- Hoffmann, J., et al. (2022). Training Compute-Optimal Large Language Models. *arXiv:2203.15556*.
- Houlsby, N., et al. (2019). Parameter-Efficient Transfer Learning for NLP. *ICML 2019*.
- Hu, E.J., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR 2022*.
- Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). Position: The Platonic Representation Hypothesis. *ICML 2024*.
- Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.
- Ilharco, G., et al. (2023). Editing Models with Task Arithmetic. *ICLR 2023*.
- Jaderberg, M., et al. (2017). Decoupled Neural Interfaces using Synthetic Gradients. *ICML 2017*.
- Jeong, J. (2026a). Bayesian BM25: A Probabilistic Framework for Hybrid Text and Vector Search. *Zenodo*. https://doi.org/10.5281/zenodo.18414940
- Jeong, J. (2026b). From Bayesian Inference to Neural Computation: The Analytical Emergence of Neural Network Structure from Probabilistic Relevance Estimation. *Zenodo*. https://doi.org/10.5281/zenodo.18512411
- Jeong, J. (2026c). Vector Scores as Likelihood Ratios: Index-Derived Bayesian Calibration for Hybrid Search. *Zenodo*. https://doi.org/10.5281/zenodo.19181568
- Jeong, J. (2026d). Answer Bandwidth: Why Sigmoid Fails in Hidden Layers. *Zenodo*. https://doi.org/10.5281/zenodo.19

254501

- Karpathy, A. (2025). nanochat: The best ChatGPT that $100 can buy. github.com/karpathy/nanochat.
- Lee, D.H., Zhang, S., Fischer, A., & Bengio, Y. (2015). Difference Target Propagation. *ECML PKDD*.
- Leviathan, Y., Kalman, M., & Matias, Y. (2023). Fast Inference from Transformers via Speculative Decoding. *ICML 2023*.
- Löwe, S., O'Connor, P., & Veeling, B. (2019). Putting An End to End-to-End: Gradient-Isolated Learning of Representations. *NeurIPS 2019*.
- Ma, S., et al. (2020). HSIC Bottleneck: Deep Learning without Back-Propagation. *AAAI 2020*.
- Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., & Rodolà, E. (2023). Relative Representations Enable Zero-Shot Latent Space Communication. *ICLR 2023*.
- Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.
- Shazeer, N., et al. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *ICLR 2017*.
- Vaswani, A., et al. (2017). Attention Is All You Need. *NeurIPS 2017*.
- Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *ACL 2019*.
- Wortsman, M., et al. (2022). Model Soups: Averaging Weights of Multiple Fine-tuned Models Improves Accuracy Without Increasing Inference Time. *ICML 2022*.