TR-Hash 500M: Deterministic Token-ID Routing for Shared Residual Experts

Boris Peyriguere
Complexity-ML, Independent Research
boris@complexity-ai.fr

August 2026

Abstract

Learned mixture-of-experts routers jointly change model capacity, parameter selection, optimization, and dispatch. We document a narrower systems design in which token identity deterministically selects two small residual experts while every token retains an always-on shared SwiGLU path. The route chooses parameters, not representations: both shared and selected branches transform the contextual hidden state produced by grouped-query attention. The realized decoder has 492,097,536 trainable parameters, 24 layers, hidden size 1,024, a 32,000-entry tokenizer, and a 2,048-token context. Each feed-forward block uses a width-4,864 shared SwiGLU and four width-128 residual experts, two active per token with fixed weights. We audit the exported checkpoint directly: all 24 route tables are distinct, every token selects two different experts, and each expert occupies exactly 16,000 of the 64,000 route slots per layer. The model completed 76,293 optimizer steps, or 19,999,752,192 FineWeb-Edu tokens. The last scheduled held-out evaluation at step 76,000 records NLL 2.661519 and perplexity 14.32. A zero-shot PIQA sanity check reaches 68.88% raw and 69.53% length-normalized accuracy. These measurements establish a completed, exportable pretraining run, not an advantage over dense or learned routing: no matched baseline, multi-seed comparison, or contamination analysis is available, and instruction post-training is outside this report.

1 Introduction and motivation

Mixture-of-experts (MoE) language models commonly learn a contextual router that selects parameters for each hidden state [3, 4, 7]. This is expressive, but it also entangles several questions. A quality change may come from additional expert capacity, a better contextual decision, router optimization, load-balancing objectives, or the dispatch implementation. Routes may drift during training and can collapse without balancing controls.

Fixed token-ID routing is a deliberately restricted alternative. It asks whether stable lexical identity can address a small residual parameter subspace while attention and a dense shared MLP continue to carry context. The route is reproducible from a persisted table, requires no learned gate, and can be audited independently of model activations. Hash Layers demonstrated that fixed lexical hashes can be useful routing signals [6]; shared-expert architectures separately motivate keeping common computation active for every token [2, 5].

This paper documents the realized 500M-class system rather than claiming an architectural win. Its contributions are:

1. an exact description of the 492.1M shared-plus-residual architecture;
2. a checkpoint-level audit of parameters and layer-specific routes;
3. a complete 20B-token pretraining trajectory with interruption-aware metric provenance; and
4. a migration contract that preserves trained routes across the training and public inference runtimes.

2 Architecture

2.1 Contextual backbone

The model is a pre-norm causal decoder with rotary position embeddings, RMSNorm, query/key normalization, tied input/output embeddings, and grouped-query attention (GQA) [1, 8]. It has 24 layers, hidden size 1,024, 16 query heads, four key/value heads, vocabulary size 32,000, and maximum position length 2,048. Deterministic routing changes only the feed-forward sublayer; it does not replace attention or contextual hidden states.

2.2 Shared and routed computation

Let $\mathbf{x}_l$ be the contextual hidden state at layer l, t a token ID, S_l the shared SwiGLU, and $E_{l,e}$ one of four residual SwiGLU experts. A fixed table $R_l \in \{0,1,2,3\}^{2\times 32000}$ supplies two routes $R_{l,1}(t) \neq R_{l,2}(t)$. The feed-forward

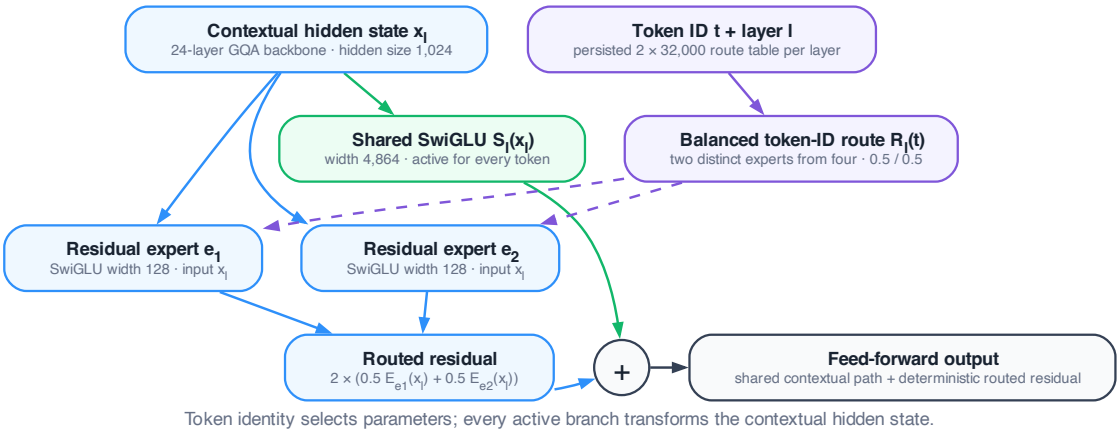

Figure 1: **Realized TR-Hash 500M feed-forward block.** GQA produces a contextual hidden state $\mathbf{x}_l$. Every token traverses a width-4,864 shared SwiGLU. A persisted layer-specific token-ID table selects two distinct width-128 residual experts from four. Both selected experts transform the same contextual state and are combined with fixed 0.5/0.5 weights and routed output scale two. The checkpoint does not enable learned channel modulation.

output is

$$F_l(\mathbf{x}_l, t) = S_l(\mathbf{x}_l) + 2 \sum_{j=1}^{2} 0.5\, E_{l,R_{l,j}(t)}(\mathbf{x}_l). \qquad (1)$$

The shared branch has width 4,864. Four stored experts each have width 128, giving stored intermediate width $4864 + 4 \times 128 = 5376$. Two experts are active, so active intermediate width is $4864 + 2 \times 128 = 5120$. The selected experts remain contextual because their inputs are $\mathbf{x}_l$, not token embeddings.

The exported configuration records shared scale one, routed scale two, top-$k = 2$, and fixed primary weights 0.5. Hashed channel modulation is disabled in this checkpoint. The narrower claim matters: the trained 500M model tests deterministic expert selection, not the later two-hash channel-modulation prototype.

2.3 Checkpoint audit

Direct inspection of the BF16 safetensors export finds 492,097,536 trainable parameters. Route tables are non-trainable buffers with shape $2 \times 32,000$ per layer. Across the exported checkpoint:

- all 24 layer route tables have different SHA-256 values;
- no token selects the same expert twice;
- each expert appears in exactly 16,000 of 64,000 route slots per layer;
- expert tensors have realized shapes $4 \times 1024 \times 128$ for gate/up and $4 \times 128 \times 1024$ for down projections.

Balance here is marginal over the two slots. We do not claim that all 12 possible ordered pairs occur equally often; the realized layer tables contain eight directed pair types.

3 Execution and export contract

Training uses `TRHashEngine`. The public checkpoint retains an equivalent legacy tensor layout for the custom inference runtime: expert tensors are a rename rather than a reshape, and the $2 \times 32,000$ route table for every layer is transplanted exactly. Regenerating routes would silently associate trained expert weights with different token sets, so conversion fails rather than guessing when a tensor is unsupported. Framework regression tests report bit-exact BF16 logits for the real 500M conversion.

The canonical engine provides a universal PyTorch reference, a general CUDA/Triton count–group–GEMM–reduce path, and a hash-native fused top-2 path for up to four experts. The optimized path compiles vocabulary routes into compact metadata, partitions tokens by expert, applies grouped expert projections, and reduces the two routed contributions. CUDA Graph buckets are available for fixed inference shapes. These are implementation capabilities, not universal speed claims; backend, hardware, precision, batch, and request shape must accompany any throughput number.

4 Pretraining record

The model was trained on FineWeb-Edu with the project 32k tokenizer. Each optimizer step contains 262,144 tokens. The final checkpoint is step 76,293, for an exact total of 19,999,752,192 tokens. The metrics record a shared/base learning-rate peak of 3×10^{-4} and an expert rate of 6×10^{-4}; the final logged values are 3×10^{-5} and 6×10^{-5}.

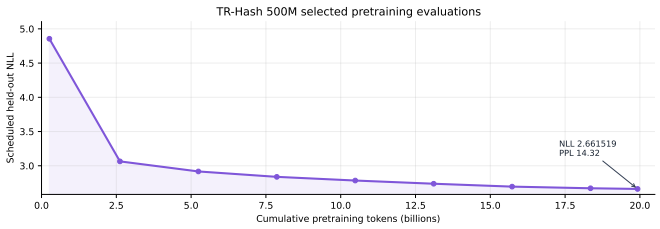

Figure 2: **Selected scheduled held-out evaluations from the 20B-token run.** The nine plotted points are copied from the interruption-aware merged metric trajectory. The last scheduled evaluation is step 76,000 (19.923B tokens), NLL 2.661519 and perplexity 14.32.

Table 1: Selected held-out evaluations.

Tokens (B)	NLL	PPL
0.262	4.855413	128.43
2.621	3.064029	21.41
5.243	2.916932	18.48
7.864	2.838020	17.08
10.486	2.784250	16.19
13.107	2.737503	15.45
15.729	2.695297	14.81
18.350	2.671209	14.46
19.923	2.661519	14.32

The raw metric file contains interrupted branches caused by repeated resumes from step 55,000. The reported merged trajectory keeps the original path through step 55,000 and the final active resume thereafter, removes overlapping abandoned rows, and yields 7,630 unique logged steps through step 76,290. Its SHA-256 is recorded in the evidence manifest. This reconstruction affects metric bookkeeping, not checkpoint weights.

As a small independent sanity check, the pretrained checkpoint was scored on all 1,838 PIQA validation examples without a chat template. Causal multiple-choice log-likelihood gives 68.88% raw accuracy and 69.53% length-normalized accuracy. No document-level decontamination against FineWeb-Edu was performed, and there is no matched dense or learned-router baseline; PIQA is therefore reported as model characterization, not comparative evidence.

5 Limitations and falsification plan

This report has one architecture, one pretraining seed, one data mixture, and no matched control. The decreasing held-out NLL shows successful language-model pretraining, but cannot attribute quality to deterministic routing. The route construction, shared width, expert width, expert learning-rate multiplier, and kernel implementation are not independently ablated. The evaluation stream is from the FineWeb-Edu source family, and no contamination audit is available. The single PIQA result is insufficient to characterize general capability.

The next credible architectural study must freeze the evaluation protocol before training and compare: (i) a parameter-matched dense model; (ii) the reported fixed token-ID routes; and (iii) a learned contextual top-2 router with the same expert tensors. Multiple seeds, external-corpus NLL, standard zero-shot tasks, contamination checks, equal-token and equal-wall-clock budgets, and route/kernel ablations are required. The existing instruction-tuning work is intentionally excluded because assistant behavior remains under evaluation.

6 Conclusion

TR-Hash 500M is a completed 492.1M-parameter language-model pretraining run in which stable token-ID tables address narrow residual experts above an always-on contextual shared MLP. Checkpoint inspection confirms the declared tensor shapes, distinct layer tables, distinct top-2 routes, and exact marginal expert balance. The model consumed almost exactly 20B tokens and reached held-out NLL 2.661519 at the last scheduled evaluation. The result establishes that the architecture, resume path, export conversion, and public runtime form a working system. It does not establish that deterministic routing is better than dense or learned routing; that question remains open for a pre-registered matched study.

Reproducibility. Implementation and tests are maintained at https://github.com/Complexity-ML/complexity-framework. The source package contains selected metric rows, hashes of the merged metrics and public checkpoint, and the route-audit summary.

Use of generative AI tools. Generative AI tools assisted with language editing, consistency checks against author-provided metrics and checkpoint tensors, figure scripting, and packaging. The author reviewed the numerical results, claims, source, and final text and takes responsibility for the work.

References

[1] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

[2] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

[3] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

[4] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[5] Huy Nguyen, Thong T. Doan, Quang Pham, Nghi D. Q. Bui, Nhat Ho, and Alessandro Rinaldo. On deepseekmoe: Statistical benefits of shared experts and normalized sigmoid gating. *arXiv preprint arXiv:2505.10860*, 2025.

[6] Stephen Roller, Sainbayar Sukhbaatar, Arthur Szlam, and Jason Weston. Hash layers for large sparse models. *Advances in Neural Information Processing Systems*, 34:17555–17566, 2021.

[7] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[8] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.