ASSOCIATIVE-RETRIEVAL REGIME ENTRY VARIES WITH HEAD PARTITION UNDER A FIXED TRAINING PROTOCOL IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Partitioning a fixed attention space across heads can be associated with different associative retrieval outcomes under a shared training protocol. We train causal transformers with six layers and $d_{\text{model}} = 512$ for 20,000 AdamW steps. Each model must recall the value paired with a queried key, using frozen sentence representations from WikiText-103 and AG News. On WikiText, we hold total attention width at 512 and move from 32 heads of dimension 16 to 4 heads of dimension 128. The fitted boundary at 50% success shifts from 5.50 to 7.41 associations. The four fitted boundary estimates are ordered across the tested partitions. A raw load selected on WikiText transfers poorly to AG News, so we map AG News loads separately. We write Hh-Dd for h heads of dimension d. At the frozen AG News operating point of six associations, H8-D64 enters the prespecified regime in 60.0% of runs. H8-D32 does so in 16.7%. An independent comparison at fixed width then matches total attention width, feed-forward width, parameter count, parameter shapes, initialization, data, and training budget. H8-D64 succeeds in 60.8% of 120 runs, versus 45.8% for H16-D32. The paired risk difference is 0.150 (95% CI [0.025, 0.275]). These results show consistent differences in entry into the associative regime across the tested head partitions under the shared training protocol.

1 INTRODUCTION

Associative retrieval near an empirically observed training transition does not always degrade smoothly. Under the same architecture and task, some training runs learn reliable key-value binding. Other runs remain near the present-value baseline. This variation raises a question that average accuracy alone cannot answer. We need to know what determines whether training reaches a successful associative-binding regime. Attention geometry provides several possible explanations for this behavior. Increasing the head dimension changes the space available within each attention head. However, this change can also alter the total width of the attention module. Conversely, increasing the head count can divide the same attention space into more, smaller subspaces. These structural quantities remain tightly coupled in many transformer configurations. Consequently, separating the individual effects of per-head dimension, total attention width, and the partition of that width across heads becomes difficult.

We study these factors using controlled associative retrieval on frozen representations derived from real text. By varying the associative load, we estimate the point at which retrieval shifts from reliable performance toward near-baseline behavior. The present-value baseline, $1/F$, changes with load. Therefore, raw load alone does not provide a fair basis for comparison across datasets. We use normalized binding to compare operating points. We treat the capacity boundary as dependent on the representation, rather than assuming that the same raw load is equally difficult for every dataset.

The experiments yield three main findings. First, under a fixed-width WikiText mapping protocol, fitted retrieval boundaries are ordered across four tested head partitions. Second, a raw load selected from WikiText places both tested architectures near the failure floor on AG News, motivating corpus-specific load remapping. Third, in an independent AG News cohort, H8-D64 shows higher regime-entry probability than H16-D32 when total attention width, feed-forward width, parameter count,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

parameter shapes, data, training budget, and same-seed initialization are matched. We interpret all three results conditionally on the shared optimization protocol.

2 Related Work

Attention geometry and memorization. Multi-head attention partitions representations into parallel query, key, and value subspaces (Vaswani et al., 2017). Prior work shows that individual heads are not interchangeable: many can be pruned with little loss (Michel et al., 2019), while others develop specialized functions (Voita et al., 2019). Head dimension can also impose a low-rank bottleneck (Bhojanapalli et al., 2020), and alternative designs such as grouped-query attention modify how attention capacity is distributed across heads (Ainslie et al., 2023). More directly, Mahdavi et al. (2024) analyze multi-head-attention memorization as head count varies, while Amsel et al. (2025) show that restricting per-head rank can limit representational power even with additional heads. Yu et al. (2026) identify settings in which larger head count can instead improve approximation. Thus the qualitative observation that smaller per-head rank can hurt is not itself new; our focus is the empirical retrieval boundary and regime-entry behavior induced by different fixed-width head partitions.

Associative memory and transformer retrieval. Associative memory originates with Hopfield networks (Hopfield, 1982) and was extended to dense memories with substantially larger storage capacity (Krotov & Hopfield, 2016; Demircigil et al., 2017). Modern Hopfield networks connect associative retrieval directly to attention (Ramsauer et al., 2021), with later work studying larger-capacity and alternative retrieval regimes (Krotov & Hopfield, 2021; Millidge et al., 2022; Burns & Fukai, 2023; Lucibello & Mézard, 2024; Hoover et al., 2024; Cabannes et al., 2024a). Transformer-specific analyses likewise study memorization and factual recall (Kajitsuka & Sato, 2025; Nichani et al., 2025), while associative recall is widely used to probe sequence models and efficient attention mechanisms (Bietti et al., 2023; Arora et al., 2024a;b). Our task uses the same key–value retrieval primitive but varies associative load to identify an empirical transition before comparing architectures near that operating region.

Learning dynamics and abrupt transitions. Whether retrieval-like behavior emerges can depend strongly on the training distribution and optimization dynamics (Garg et al., 2022; Chan et al., 2022; Cabannes et al., 2024b). Abrupt formation of induction-style retrieval circuits has also been linked to in-context learning (Olsson et al., 2022; Singh et al., 2024), while related work studies abrupt learning under changing data statistics and optimization conditions (Reddy, 2024; Nanda et al., 2023). Because apparently sharp transitions can depend on the evaluation metric (Wei et al., 2022; Schaeffer et al., 2023), we report continuous held-out accuracy and normalized binding in addition to the binary regime-entry endpoint. We use *regime* descriptively, not as a claim of a formal phase transition.

3 Experimental Framework

3.1 Task, Data, and Implementation

We study associative retrieval using frozen text representations from WikiText-103 (Merity et al., 2017) and AG News (Zhang et al., 2015). Text is encoded with the frozen `all-MiniLM-L6-v2` sentence encoder (Reimers & Gurevych, 2019; Wang et al., 2020), which produces normalized representations with 384 dimensions. We do not use the AG News category labels. The corpus serves only as an independent distribution of text representations.

For WikiText, we treat each raw text record in the dataset as one candidate text unit after whitespace normalization. No additional sentence tokenizer is applied. We discard empty records and WikiText headings, and we keep units of 8 to 40 words, counted by whitespace. We then remove exact overlaps across the official splits. The result is 50,000 training, 258 validation, and 252 test units. For AG News, we keep texts of 8 to 80 words and split the official training set with a deterministic hash. This gives 50,000 training and 5,000 validation texts. We also keep 5,000 examples from the official test split. Exact duplicate texts are removed across train, validation, and test. Each query is a deterministic view of its key in which 10% of the words are dropped. Each view keeps at least six

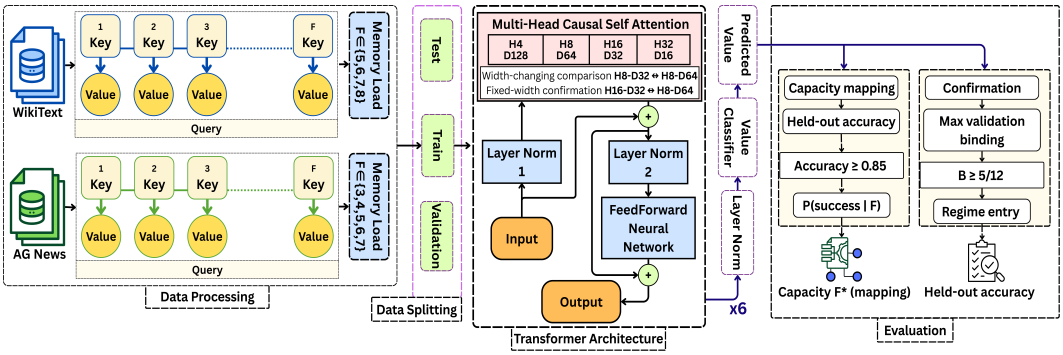


Figure 1: Experimental framework for capacity mapping, load selection, and independent confirmation.

words, and texts with six or fewer words are left unchanged. A learned linear projection without bias maps the frozen sentence representation from 384 to 512 dimensions, the width of the transformer residual stream.

Each example contains F key–value associations,

$$\mathcal{M} = \{(k_1, v_1), \ldots, (k_F, v_F)\},$$

followed by a query q matching one stored key. The model must return the corresponding value. Each episode contains F distinct value identities drawn from a 128-class vocabulary. Random guessing over the full output space therefore has accuracy $1/128$. We additionally use $1/F$ as the *present-value baseline*: the accuracy of randomly selecting among the F values that appear in the episode without identifying the correct key–value binding.

The associative learner is a transformer with six layers and $d_{\text{model}} = 512$. We denote an architecture with H attention heads and per-head dimension d_h as HH-Dd_h. For example, H8-D64 has $H = 8$ and $d_h = 64$. Episodes are serialized as $[k_1, v_1, \ldots, k_F, v_F, q]$. Values are drawn without replacement from the 128-class vocabulary. The WikiText capacity map uses $d_{\text{ff}} = 2048$. Models are trained with AdamW for 20000 steps at batch size 256. This optimization uses a learning rate of 10^{-4} and a weight decay of 10^{-4}. We use 500 warmup steps followed by cosine decay to zero at step 20000. Training and evaluation use fp16 autocast with gradient scaling.

The transformer uses pre-layer normalization and learned positional embeddings. It incorporates GELU feed-forward blocks and bias-free Q/K/V/O and FFN projections with no attention dropout. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.98)$ and gradient clipping at 1.0. We use mixed-precision training. Validation is logged every 1000 steps on 4096 fixed episodes. The final 20000-step model is evaluated on 4096 fixed held-out test episodes. Value identities use a learned 128-by-512 embedding matrix shared with the output classifier.

Across the experimental sequence we compare architectures that vary head count, per-head dimension, or their product. Dedicated fixed-width comparisons hold $d_{\text{attn}} = Hd_h$ constant. In the later parameter-matched comparisons, FFN width is adjusted to offset changes in total attention width. This adjustment ensures that the compared architectures have equal numbers of trainable parameters. The fixed-width H8-D64 versus H16-D32 control uses $d_{\text{ff}} = 3072$ for both architectures. The synthetic experiments were used only during the discovery stage and do not provide the main evidence for the final claims; they are reported in the appendix. Figure 1 summarizes the real-text task and experimental sequence.

3.2 Endpoints and Capacity Definitions

Let A be retrieval accuracy at associative load F. Normalized binding is defined as

$$B(A, F) = \frac{A - \frac{1}{F}}{1 - \frac{1}{F}}. \tag{1}$$

Thus, $B = 0$ corresponds to the present-value baseline $1/F$, and $B = 1$ to perfect retrieval.

A run enters the successful binding regime in the confirmatory experiments when

$$B_{\max,\mathrm{val}} \geq \tau, \qquad \tau = \frac{5}{12}.$$

This corresponds to raw validation accuracies of 0.500 at $F = 7$ and 0.5139 at $F = 6$.

The normalized threshold $\tau = 5/12$ was used for the WikiText confirmation. It is not treated as preregistered for that experiment, but was frozen unchanged before the subsequent AG News mapping and confirmatory cohorts. The WikiText threshold is not treated as preregistered; it was subsequently frozen unchanged before the AG News mapping and confirmatory cohorts. The earlier WikiText capacity map used the separate criterion $A_{\mathrm{test}} \geq 0.85$. Throughout this paper, "capacity boundary" denotes this empirical 50%-success boundary under the specified training and evaluation protocol. It is not intended as an optimizer-independent representational limit.

3.3 Architectural Comparisons

The study contains two distinct comparison families. **First,** the WikiText capacity map holds total attention width fixed at $d_{\mathrm{attn}} = 512$ while varying its partition: H32-D16, H16-D32, H8-D64, and H4-D128. This experiment tests whether different partitions of the same attention space shift the empirical capacity boundary. **Second,** the H8-D64 versus H8-D32 comparison holds head count fixed but changes both head dimension and total attention width (512 versus 256). We use this comparison to study regime accessibility and cross-corpus load transfer, not to isolate head partition. **The final** AG News confirmation returns to the fixed-width question by comparing H8-D64 with H16-D32. Both use $d_{\mathrm{attn}} = 512$, $d_{\mathrm{ff}} = 3072$, identical parameter tensor shapes, and byte-identical same-seed initialization. Changing the standard head partition also changes the usual $1/\sqrt{d_h}$ attention-logit scaling. We therefore interpret this experiment as an intervention on the standard head partition as a whole, not as a temperature-decoupled effect of dimensionality alone.

3.4 Experimental Sequence

Capacity mapping, exploration, and confirmation use separate seed cohorts; Table 1 summarizes the sequence. WikiText mapping selects $F = 7$ for the first independent comparison. Because the same load places both architectures near the floor on AG News, AG News is mapped separately over $F \in \{3, 4, 5, 6, 7\}$. For WikiText, $F = 7$ is the nearest tested integer load to the H8-D64 fitted mapping boundary, $F^\star = 6.808$; H8-D32 was subsequently evaluated at that same frozen load. This architecture-specific operating-point rule predates, and is distinct from, the pooled rule used later for AG News.

The confirmation load is chosen by

$$F_{\mathrm{confirm}} = \arg\min_{F} \left| \hat{p}_{\mathrm{pooled}}(F) - 0.5 \right|, \tag{2}$$

with ties resolved toward the smaller F. The rule uses pooled difficulty, not the architecture difference, and selects $F = 6$. Confirmation then uses 120 new paired seeds, followed by a separate 120-pair fixed-width control.

The 120-pair AG News confirmation size was fixed before outcome analysis. Under the planning probabilities $p_{\mathrm{wide-only}} = 0.30$ and $p_{\mathrm{control-only}} = 0.10$ inherited from the mapping stage, the exact unconditional McNemar power is 0.934 at $\alpha = 0.05$.

3.5 Statistical Analysis

The primary confirmatory outcome is paired binary regime entry. For paired seed s, let $Y_s^{(w)}$ and $Y_s^{(c)}$ denote binary outcomes for the wider-head configuration and its control. The paired risk difference is

$$\widehat{\Delta}_{\mathrm{RD}} = \frac{1}{N} \sum_{s=1}^{N} \left(Y_s^{(w)} - Y_s^{(c)} \right). \tag{3}$$

Confidence intervals for paired risk differences are obtained by paired percentile-bootstrap resampling, and paired binary outcomes are tested with the exact two-sided McNemar test. Because three

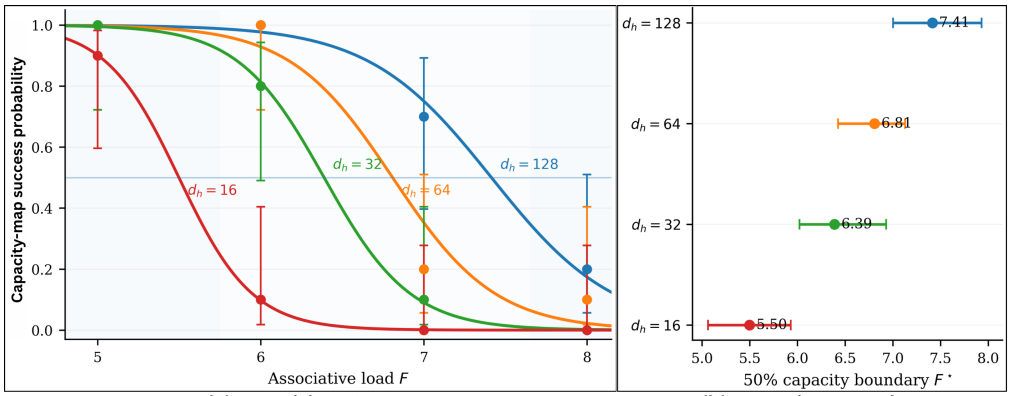


Figure 2: WikiText fixed-width capacity map. (a) Success transitions across load; (b) fitted 50% capacity boundaries with 95% bootstrap intervals.

confirmatory McNemar comparisons are reported, we also apply a Holm family-wise adjustment across these tests as a post-hoc sensitivity analysis. The adjusted p values are 1.95×10^{-3} for WikiText, 9.20×10^{-12} for the AG News confirmation, and 0.0328 for the fixed-width confirmation. Held-out test accuracy is the primary continuous secondary outcome; its paired mean difference is reported with a percentile-bootstrap 95% confidence interval. The two 120-pair AG News cohorts use 50,000 paired bootstrap resamples.

Normalized held-out binding is an affine transformation of accuracy at fixed F and is therefore reported for scale comparability rather than treated as an independent hypothesis. Maximum validation binding is reported descriptively. Held-out binary success is evaluated separately from the validation-defined endpoint; in both 120-pair AG News cohorts, the marginal success counts agree with the validation-defined counts.

Matched seeds are used for the architecture comparisons. Mapping seeds are not reused for confirmation, and the AG News confirmation load is selected without access to the later confirmatory cohorts. The AG News threshold, load-selection rule, and sample size were frozen before the corresponding confirmatory runs.

Table 1: Main mapping and confirmation sequence.

Study	**Models**	F	d_{attn}	d_{ff}	**Parameters**	N
WikiText capacity map	H32-D16 / H16-D32 / H8-D64 / H4-D128	5–8	512	2048	19,158,528	10/point
WikiText confirmation	H8-D32 / H8-D64	7	256 / 512	3584 / 3072	25,448,960	40 pairs
AG News raw transfer	H8-D32 / H8-D64	7	256 / 512	3584 / 3072	25,448,960	30 pairs
AG News load map	H8-D32 / H8-D64	3–7	256 / 512	3584 / 3072	25,448,960	10/load
AG News confirmation	H8-D32 / H8-D64	6	256 / 512	3584 / 3072	25,447,936	120 pairs
AG News fixed-width	H16-D32 / H8-D64	6	512 / 512	3072 / 3072	25,447,936	120 pairs

All real-text cohorts use $d_{\text{model}} = 512$ and six transformer layers. Q/K/V/O and FFN linear projections are bias-free. The WikiText capacity map uses $A_{\text{test}} \geq 0.85$; subsequent cohorts use $B_{\text{max,val}} \geq 5/12$. N denotes seeds or matched seed pairs.

4 WIKITEXT FIXED-WIDTH CAPACITY MAP AND EXPLORATORY CONTROLS

The WikiText experiments first test whether head geometry changes the location of the associative-binding transition. Head dimension is varied across associative loads, and the resulting capacity boundary is then used to choose an operating point for independent confirmation.

4.1 CAPACITY MAPPING AT FIXED ATTENTION WIDTH

Figure 2 maps the WikiText transition while holding total attention width fixed at $d_{\text{attn}} = 512$. The four configurations are H32-D16, H16-D32, H8-D64, and H4-D128. Thus head count and head dimension co-vary while their product remains constant.

For each partition we fit

$$P(\text{success} \mid F) = \sigma(k(F^\star - F))$$

by bounded binomial-likelihood optimization using L-BFGS-B, with $F^\star \in [F_{\min} - 5, F_{\max} + 5]$ and $k \in [10^{-3}, 100]$. Completely separated resamples therefore return finite boundary solutions under the bounded fit. Confidence intervals use 5,000 within-load seed-resampling refits.

The fitted capacity boundaries are

$$F^\star = 5.50,\ 6.39,\ 6.81,\ 7.41$$

for $d_h = 16, 32, 64, 128$, respectively. The corresponding 95% confidence intervals are shown in Figure 2(b).

The difference is concentrated near the transition. At low loads the models are close to the ceiling, while above the boundary they approach the failure floor. The fitted success curves change most strongly near the estimated boundaries, motivating independent comparisons at loads in this region.

As a post-hoc analysis of held-out test accuracy in the completed mapping cohort, we tested the ordered alternative $d_h = 16 < 32 < 64 < 128$ using Page's repeated-measures trend test on matched seeds. The statistics are $L = 298, 295, 292, 276$ for $F = 5, 6, 7, 8$, respectively, with Holm-adjusted one-sided p values 2.81×10^{-11}, 5.88×10^{-9}, 1.64×10^{-7}, and 0.00205. Across 10,000 paired seed-resampling draws, the four fitted boundaries retain the strict ordering in 92.27% of draws. The observed extreme-partition difference is $\Delta F^\star = 1.916$ (95% paired-bootstrap CI $[1.126, 2.616]$).

4.2 EXPLORATORY ALTERNATIVE EXPLANATIONS

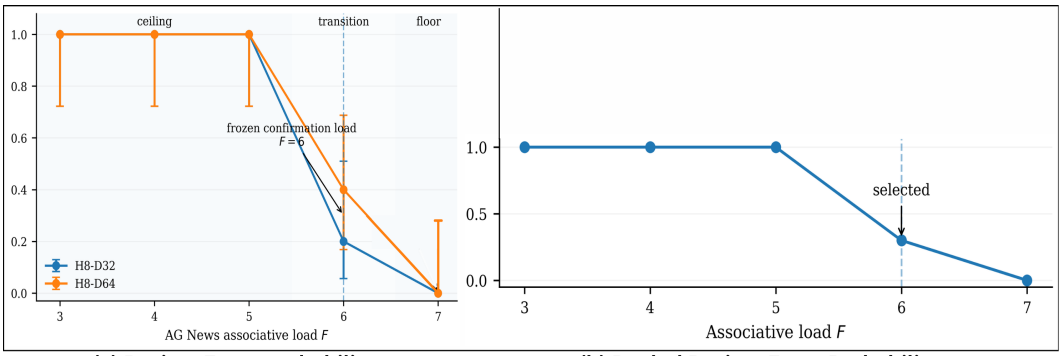
Two exploratory interventions do not support a simple one-dimensional explanation. At $F = 7$, varying active Q/K dimension from $16, 32, 64, 128$ with value-head dimension fixed at 64 gives regime-entry rates $0.70, 0.40, 0.30, 0.60$, respectively (Appendix Table 5), a non-monotonic pattern. In the parameter-matched WikiText factorial, the width, head-count, and interaction effects on final test accuracy are 0.101, -0.053, and -0.164, respectively, with all 95% confidence intervals spanning zero (Appendix Table 6).

5 CROSS-CORPUS LOAD REMAPPING AND INDEPENDENT CONFIRMATION

The H8-D32 versus H8-D64 comparisons in this section change both per-head dimension and total attention width, with FFN width adjusted for parameter matching; they therefore characterize transfer and regime accessibility but do not isolate head partition. The WikiText result is first tested with new seeds, then transferred to a second representation distribution. The AG News experiments also test whether the same raw associative load represents the same difficulty across corpora.

5.1 INDEPENDENT WIKITEXT CONFIRMATION

The WikiText capacity map places H8-D64 near its fitted transition at $F = 7$; H8-D32 is then evaluated at the same frozen load. This operating point was inherited from the $d_{\text{ff}} = 2048$ mapping



(a) Regime Entry probability (b) Pooled Regime Entry Probability

Figure 3: AG News capacity map and frozen $F = 6$ confirmation-load selection.

models; the subsequent parameter-matched confirmation uses different FFN widths and therefore does not constitute a remapping of those models' own capacity boundaries.

At this load, 40 new paired seeds compare H8-D64 with H8-D32. H8-D64 enters the successful regime in 35.0% of runs, compared with 2.5% for H8-D32. The paired risk difference is 0.325 (95% CI $[0.175, 0.500]$), with exact two-sided McNemar $p = 9.77 \times 10^{-4}$. The difference also appears in held-out performance: mean test accuracy is 0.3999 for H8-D64 and approximately 0.162 for H8-D32. Thus, the result is not limited to crossing the validation threshold.

5.2 The Same Raw Load Does Not Transfer to AG News

The first AG News experiment uses the same raw load, $F = 7$. Across 30 paired seeds, H8-D64 succeeds in only $1/30$ runs, while H8-D32 succeeds in $0/30$. Both architectures are therefore operating near the failure floor. This failed transfer is consistent with the same raw load F having different effective difficulty under the two corpus-specific representation and preprocessing pipelines. Consequently, AG News requires its own capacity map before an independent architecture comparison can be made near the transition region.

5.3 Mapping the AG News Transition

AG News is mapped over $F \in \{3, 4, 5, 6, 7\}$ using 10 paired seeds per architecture and load. Figure 3 shows the resulting transition. For $F = 3, 4, 5$, both architectures succeed in 10/10 runs; at $F = 6$, H8-D32 succeeds in 2/10 and H8-D64 in 4/10; at $F = 7$, both succeed in 0/10. Thus $F = 6$ is the only non-degenerate tested load and its pooled success rate, $6/20 = 0.30$, is also the closest tested value to the prespecified target of 0.5. The frozen rule therefore selects $F = 6$.

5.4 Independent AG News Confirmation

We then evaluate the frozen $F = 6$ comparison on 120 new paired seeds. H8-D64 enters the successful regime in 60.0% of runs. H8-D32 does so in only 16.7%. The paired risk difference is

$$\widehat{\Delta}_{\mathrm{RD}} = 0.433, \qquad 95\%\ \mathrm{CI} = [0.333, 0.533], \tag{4}$$

with exact McNemar $p = 3.07 \times 10^{-12}$.

Continuous performance points the same way. The paired held-out test-accuracy difference is 0.347 (95% CI $[0.276, 0.416]$). The equivalent normalized-binding effect is reported in Appendix Table 9. The held-out binary endpoint gives the same success counts as the endpoint defined on validation data, 72 versus 20. Across WikiText and AG News, the architectures fall in the same order at operating points that were frozen independently for each corpus. On AG News, we also mapped both tested architectures before the confirmation run. The next section asks whether total attention width alone can explain this difference.

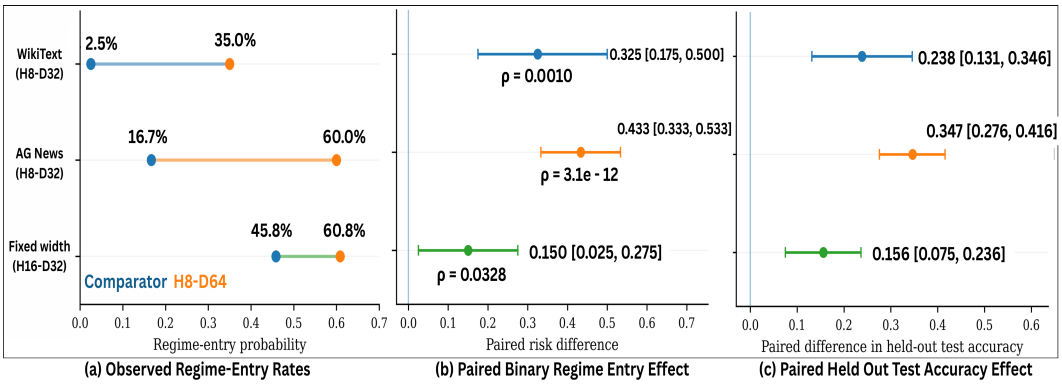


Figure 4: Confirmatory effects of attention geometry: (a) regime-entry rates, (b) paired risk differences, and (c) held-out test-accuracy differences.

6 Independent Fixed-Width Confirmation on AG News

We next test the fixed-width question on an independent AG News cohort. H8-D64 and H16-D32 both use $d_{\text{attn}} = 512$ and are otherwise matched as specified in Section 3.3.

6.1 Fixed-Width Comparison

This control compares H8-D64 with H16-D32, as defined in Section 3.3. Both models have $d_{\text{attn}} = 512$ and are matched in FFN width, parameter count, parameter shapes, data, training budget, and same-seed initialization. We ran 120 new paired seeds at the frozen AG News load $F = 6$. We retained this operating point rather than remapping the load for H16-D32. The load had already been selected for the corpus before the fixed-width cohort was run. Keeping the same value therefore avoids selecting a different operating point after observing results for one architecture.

6.2 Regime Entry at Fixed Attention Width

H8-D64 enters the successful binding regime in $73/120$ runs (60.8%). H16-D32 does so in $55/120$ (45.8%). This rate also shows that $F = 6$ sits at neither the ceiling nor the failure floor for H16-D32, although we did not estimate its full capacity curve. The paired risk difference is 0.150 (95% CI $[0.025, 0.275]$), with exact McNemar $p = 0.0328$. Of the 120 pairs, 41 succeed only for H8-D64 and 23 only for H16-D32. Another 32 succeed for both, and 24 for neither. The difference remains after matching total attention width, FFN width, parameter count, tensor shapes, initialization, data, and training budget.

6.3 Continuous Outcomes

Continuous outcomes follow the same ordering. Mean held-out test accuracy is 0.641 for H8-D64 and 0.485 for H16-D32. The paired held-out test-accuracy difference is 0.156 (95% CI $[0.075, 0.236]$). The equivalent normalized-binding effect and additional continuous-outcome statistics are reported in Appendix Table 9. The held-out binary endpoint gives the same success counts as the endpoint defined on validation data. The fixed-width comparison retains the same architecture ordering in both binary and continuous outcomes. Figure 4 summarizes the binary and continuous effects across the three confirmatory comparisons.

7 Discussion

The strongest evidence at fixed width comes from two complementary results. The first is the WikiText capacity map, which shows an ordered shift in the empirical retrieval boundary across four partitions of the same attention width. The second is the independent AG News confirmation with

120 pairs. There, H8-D64 retains its advantage over H16-D32 when attention width, FFN width, parameter count, initialization, data, and training budget are matched.

The experiments across corpora address a separate question. A raw load near the WikiText transition places both tested architectures near the failure floor on AG News. A load mapped independently on AG News shows substantial separation on new seeds. In this setup, raw associative load does not measure difficulty independently of the dataset.

All comparisons on real text use the same optimization protocol. The results should therefore be read as architectural effects conditional on that protocol, rather than as estimates of the optimal performance of each architecture. Future work can test whether the same ordering persists under broader settings of optimization and model scale.

8 Conclusion

Under a shared training protocol with 20000 steps, different partitions of a fixed attention width are associated with different associative-retrieval boundaries and regime-entry rates. The WikiText map shows an ordered boundary shift across four fixed-width partitions. An independent AG News cohort retains an H8-D64 advantage over H16-D32. This advantage holds when attention width, FFN width, parameter count, initialization, data, and training budget are matched. Separate experiments further show that raw associative load does not transfer directly across the two corpus pipelines. These findings support a consistent protocol-conditional head-partition effect in the tested associative-retrieval setting.

Reproducibility Statement

The main text specifies the associative-retrieval task, data construction, architectural comparisons, endpoint definitions, load-selection procedures, and statistical protocol. The appendix reports synthetic discovery experiments, architecture and optimization diagnostics, detailed capacity estimates, confirmatory outcomes, and robustness analyses. An anonymized repository provides the experiment code, analysis scripts, training configurations, seed ranges, provenance documentation, and available result artifacts used to reproduce the reported analyses and figures: `https://anonymous.4open.science/r/TRANSFORMERS_ASSOCIATIVE_MEMORY_Regulation/README.md`.

AI Use Statement

Large language models (LLM) were used only as a basic drafting assistant to help clean up the text, structure the formatting, and generate initial code snippets and latex formatting. The tool did not produce any data or scientific findings. The authors verified all results, code, and text, and take full responsibility for the entire study.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298/.

Noah Amsel, Gilad Yehudai, and Joan Bruna. Quality over quantity in attention layers: When adding more heads hurts. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9c537882044c8b5352c363e840872ddb-Abstract-Conference.html.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Y. Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language

models. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=LY3ukUANko.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1763–1840. PMLR, 2024b. URL https://proceedings.mlr.press/v235/arora24a.html.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 864–873. PMLR, 2020. URL https://proceedings.mlr.press/v119/bhojanapalli20a.html.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1560–1588, 2023. doi: 10.52202/075280-0077. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0561738a239a995c8cd2ef0e50cfa4fd-Abstract-Conference.html.

Thomas F. Burns and Tomoki Fukai. Simplicial Hopfield networks. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_QLsH8gatwx.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=Tzh6xAJSll.

Vivien Cabannes, Berfin Simsek, and Alberto Bietti. Learning associative memories with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 5114–5134. PMLR, 2024b.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891, 2022. doi: 10.52202/068431-1371. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2): 288–299, 2017. doi: 10.1007/s10955-017-1806-y. URL https://doi.org/10.1007/s10955-017-1806-y.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022. doi: 10.52202/068431-2217. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Benjamin Hoover, Duen Horng Chau, Hendrik Strobelt, Parikshit Ram, and Dmitry Krotov. Dense associative memory through the lens of random features. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0742. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/29ff36c8fbed10819b2e50267862a52a-Abstract-Conference.html.

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL https://doi.org/10.1073/pnas.79.8.2554.

Tokio Kajitsuka and Issei Sato. On the optimal memorization capacity of Transformers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a6fb8b1284a412543d853bfdddfc8f8e-Abstract-Conference.html.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, volume 29, pp. 1172–1180, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/eaae339c4d89fc102edd9dbdb6a28915-Abstract.html.

Dmitry Krotov and John J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=X4y_100X-hX.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7):077301, 2024. doi: 10.1103/PhysRevLett.132.077301. URL https://doi.org/10.1103/PhysRevLett.132.077301.

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Memorization capacity of multi-head attention in Transformers. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d8542126cd3e0dd6c0a44e0aa1957072-Abstract-Conference.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *Proceedings of the 5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15561–15583. PMLR, 2022. URL https://proceedings.mlr.press/v162/millidge22a.html.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Eshaan Nichani, Jason D. Lee, and Alberto Bietti. Understanding factual recall in Transformers via associative memories. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0bf9f909c24c8879d1b7f86fa50a9e49-Abstract-Conference.html.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html.

Aaditya K. Singh, Ted Moskovitz, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45637–45662. PMLR, 2024. URL https://proceedings.mlr.press/v235/singh24c.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

Penghao Yu, Haotian Jiang, Zeyu Bao, Ruoxi Yu, and Qianxiao Li. The effect of attention head count on Transformer approximation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7d72a514fc6948515af3ee69c1059776-Abstract-Conference.html.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

A Additional Experimental Results

The main paper uses the WikiText and AG News experiments as the primary evidence. This appendix reports synthetic discovery experiments, supporting diagnostics, detailed confirmatory results, and robustness analyses. Synthetic experiments are treated as exploratory evidence rather than as independent confirmation of the final real-text findings.

A.1 Synthetic Discovery Experiments

Synthetic task and protocol. Each synthetic episode contains F key–value pairs followed by a query key drawn from the presented set. Values are sampled from an eight-class vocabulary, key and value token ranges are disjoint, and the model predicts the queried value from the final-position hidden state. The synthetic learner is a two-layer decoder-only transformer with causal masking, pre-layer normalization, learned positional embeddings, and a GELU FFN of width $4d_{\text{model}}$. Attention and FFN projections use no bias. Training uses AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay 10^{-4}), linear warmup followed by cosine decay, gradient clipping at 1.0, mixed precision, and batch size 512. The fixed-$d_{\text{model}} = 128$ head-partition experiment uses learning rate 5×10^{-5} and a 20,000-step budget. Synthetic capacity success is defined by held-out accuracy at least 0.85; $F^\star$ is the fitted load at which success probability is 0.5.

The synthetic experiments are discovery evidence only and use a different task, model depth, and representation distribution from the real-text studies; their absolute $F^\star$ values should therefore not be compared numerically with the real-text boundaries.

The initial synthetic experiments varied residual width while keeping $d_h = 16$. Table 2 summarizes the resulting capacity estimates.

The normalized boundary $F^\star/d_{\text{model}}$ decreases as residual width increases, ruling out the initial hypothesis of a width-independent critical density. Because residual width and head count change together in this sweep, it was used to motivate a more controlled head-partition test. Absolute $F^\star$ is also non-monotonic in this discovery sweep, peaking near $d_{\text{model}} = 128$ and decreasing at the larger tested widths. Given the optimization sensitivity documented below, we do not interpret this sweep as a representational scaling law.

The subsequent experiment fixes $d_{\text{model}} = 128$ and varies only the number and width of attention heads. All four configurations contain $436{,}992$ trainable parameters and use the same training protocol. Boundary regions were strengthened with additional seeds and 2,000 bootstrap fits.

Figure 5 shows the resulting capacity boundaries, with exact estimates in Table 3.

The trend is monotonic over the four tested partitions. This synthetic result motivated the later real-text experiments but is not used to claim that head dimension alone determines capacity.

Table 2: Synthetic width sweep with $d_h = 16$.

d_{model}	Heads	$F^\star$	$F^\star / d_{\text{model}}$
32	2	$\approx$ 14–15	$\approx$ 0.44–0.47
64	4	15.50	0.242
128	8	$\approx$ 23.84	$\approx$ 0.186
192	12	$\approx$ 22.5	$\approx$ 0.117
256	16	$\approx$ 21–23	$\approx$ 0.082–0.090

Note. These are discovery-stage estimates from the original width sweep. The $d_{\text{model}} = 128$ configuration was later re-estimated using the dedicated head-partition experiment below.

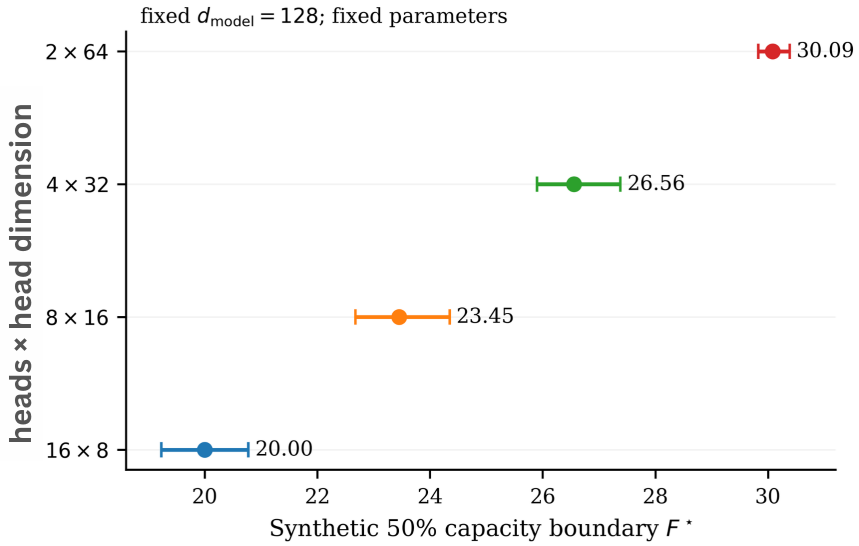


Figure 5: Synthetic capacity under fixed model width

Table 3: Synthetic fixed-width head-partition capacity estimates.

Heads	d_h	$F^\star$	95% CI
2	64	30.086	[29.829, 30.386]
4	32	26.556	[25.900, 27.380]
8	16	23.449	[22.677, 24.349]
16	8	20.000	[19.226, 20.774]

A.2 Supporting Diagnostics and Alternative Explanations

Optimization sensitivity. Two diagnostic checks showed that optimization settings can strongly change the apparent capacity boundary. Table 4 summarizes these controls.

These checks motivated the matched optimization settings used in the later capacity and confirmatory experiments.

Active Q/K dimension. To test whether larger heads help simply by providing a larger active query/key space, the value-head dimension was fixed at 64 while the active Q/K dimension was varied. Table 5 shows the resulting regime-entry probabilities.

Success probabilities are 0.70, 0.40, 0.30, and 0.60 for $d_{QK} = 16, 32, 64, 128$, respectively. The pattern is non-monotonic and therefore provides no evidence that increasing active Q/K dimensionality alone explains the main fixed-width capacity effect. We treat this experiment as a negative mechanistic intervention.

Factorial architecture analysis. A WikiText factorial at $F = 7$ compares H8-D32, H8-D64, H16-D32, and H16-D64 using 10 matched seeds per configuration. FFN width is adjusted to keep the trainable parameter count equal. Table 6 summarizes the outcomes.

The paired width main effect is 0.101 with 95% CI $[-0.063, 0.273]$; the head-count main effect is -0.053 with 95% CI $[-0.173, 0.068]$; and the width-by-head interaction is -0.164 with 95% CI $[-0.414, 0.069]$. None supports a simple architecture-wide main effect. The matched FFN widths are 3584, 3072, 3072, and 2048 for H8-D32, H8-D64, H16-D32, and H16-D64, respectively, giving 25,448,960 trainable parameters per model.

Parameter-count detail. The $F \leq 7$ parameter-matched models use a maximum learned positional table of length 15, whereas the $F = 6$-only confirmation models require length 13. The two missing 512-dimensional positional vectors account for the $1{,}024$-parameter difference between $25{,}448{,}960$ and $25{,}447{,}936$.

Table 4: Optimization controls from the synthetic experiments.

Control	Setting	Successes	Rate
Learning rate	$d_{\text{model}} = 128,\ F = 15,\ \eta = 3 \times 10^{-4}$	0/5	0.00
	$d_{\text{model}} = 128,\ F = 15,\ \eta = 5 \times 10^{-5}$	5/5	1.00
Batch size	$d_{\text{model}} = 64,\ F = 17,\ \text{BS} = 512$	3/20	0.15
	$d_{\text{model}} = 64,\ F = 17,\ \text{BS} = 4096$	20/20	1.00

Table 5: Active-Q/K intervention on WikiText at $F = 7$ with $H = 8$ and value-head dimension 64; each condition uses 10 matched seeds.

Active d_{QK}	Successful runs	Success probability
16	7/10	0.70
32	4/10	0.40
64	3/10	0.30
128	6/10	0.60

Table 6: Exploratory parameter-matched factorial on WikiText representations at $F = 7$.

Architecture	Mean final test accuracy	Regime-entry probability
H8-D32	0.191	0.10
H8-D64	0.373	0.30
H16-D32	0.219	0.10
H16-D64	0.238	0.10

Table 7: Paired binary outcome decomposition for the three confirmatory comparisons.

Study	**Both**	**Wide only**	**Control only**	**Neither**
WikiText	0	14	1	25
AG News	15	57	5	43
Fixed width	32	41	23	24

A.3 Detailed Real-Text Results

WikiText capacity estimates. Table 8 reports the fitted WikiText capacity boundaries shown graphically in the main paper.

Confirmatory comparisons. Table 9 reports the binary and continuous effects from the WikiText confirmation, AG News confirmation, and fixed-width control.

Direct AG News transfer. Before mapping the AG News transition, the WikiText operating point $F = 7$ was transferred directly. H8-D64 succeeded in only $1/30$ runs and H8-D32 in $0/30$, placing both architectures near the failure floor. This negative result motivated the separate AG News capacity map and the frozen $F = 6$ confirmation load. The paired binary outcome decompositions underlying the exact McNemar tests are given in Table 7.

Table 8: Estimated 50% capacity boundaries on the WikiText representation bank.

Heads	d_h	$F^\star$	95% CI
4	128	7.414	[7.002, 7.931]
8	64	6.808	[6.426, 7.130]
16	32	6.388	[6.021, 6.931]
32	16	5.498	[5.064, 5.929]

All observed fits and all 5,000 bootstrap refits returned finite solutions under the bounded likelihood fit.

Table 9: Confirmatory results across WikiText, AG News, and the fixed-width control.

Study	**Regime entry**	**Risk diff.**	**Δ test acc.**	**Δ test binding**
WikiText	35.0% vs. 2.5%	0.325 [0.175, 0.500]	0.238 [0.131, 0.346]	0.278 [0.155, 0.405]
AG News	60.0% vs. 16.7%	0.433 [0.333, 0.533]	0.347 [0.276, 0.416]	0.416 [0.332, 0.500]
Fixed width	60.8% vs. 45.8%	0.150 [0.025, 0.275]	0.156 [0.075, 0.236]	0.187 [0.089, 0.284]

WikiText and AG News compare H8-D64 against H8-D32. The fixed-width control compares H8-D64 against H16-D32. Exact McNemar p values are 9.77×10^{-4}, 3.07×10^{-12}, and 0.0328 for WikiText, AG News, and the fixed-width control, respectively. As a post-hoc sensitivity analysis, a Holm family-wise adjustment across the three confirmatory McNemar tests gives adjusted p values of 1.95×10^{-3} for WikiText, 9.20×10^{-12} for the AG News confirmation, and 0.0328 for the fixed-width confirmation.

A.4 Seed-Level Regime Structure

Figure 6 shows held-out normalized binding for all runs in the two 120-pair AG News cohorts. Both comparisons exhibit substantial mass near the present-value baseline and at substantially higher binding values, making the seed-level outcome structure visible without relying only on the binary regime-entry endpoint. These distributions are descriptive; the primary inference remains the paired validation-defined regime-entry comparison.

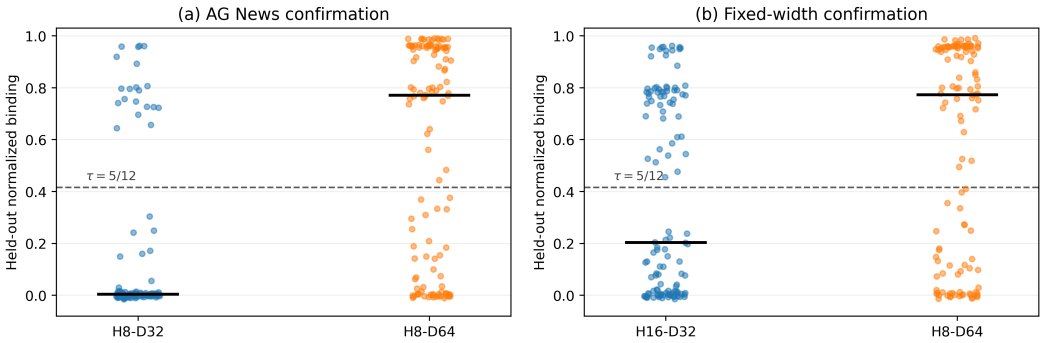
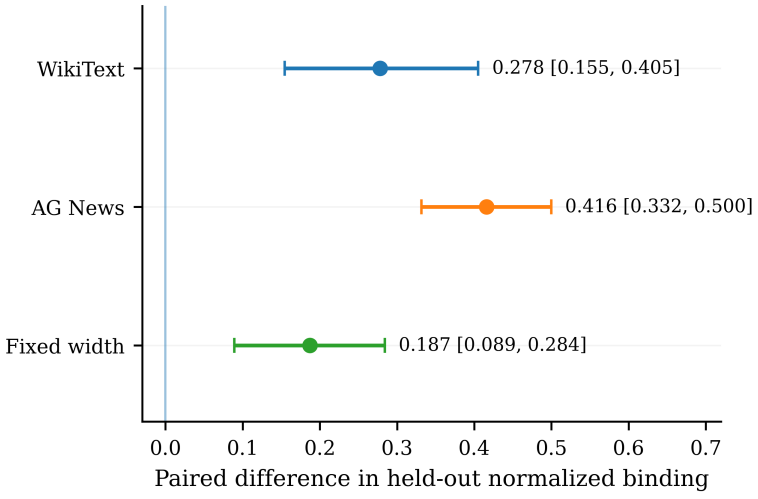


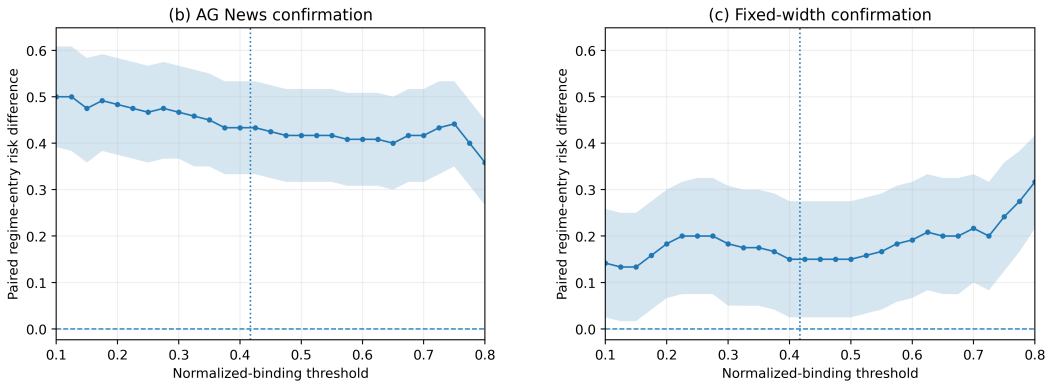
Figure 6: Seed-level held-out normalized binding in the two 120-pair AG News cohorts. Points denote individual runs; black bars mark medians and the dashed line marks $\tau = 5/12$.

A.5 Robustness of the Confirmatory Effects

The confirmatory ordering is also present in continuous held-out binding. We additionally vary the normalized-binding threshold to test whether the binary conclusions depend on the prespecified $\tau = 5/12$ cutoff. Figure 7 summarizes these checks; shaded regions show 95% paired-bootstrap confidence intervals.



(a) Detailed confirmatory effects.



(b) AG News confirmation.

(c) Fixed-width control.

Figure 7: Robustness checks. (a) Held-out binding effects; (b–c) paired regime-entry risk differences across alternative normalized-binding thresholds. Vertical dotted lines mark $\tau = 5/12$; shaded regions are pointwise 95% paired-bootstrap intervals.