Distilled Reasoning Reshapes the Linear Predictability of fMRI Naturalistic-Listening BOLD

Anonymous ACL submission

Abstract

Reasoning-trained language models now dominate practical NLP, yet whether they exhibit a brain-encoding signature distinct from base or instruction-tuned models is unknown. This matters: brain-encoding scores are an increasingly common tool to assess model quality, and an undetected distillation effect would bias both such evaluations and our understanding of what distillation changes inside the model. We compare seven models on naturalistic-listening fMRI from two independent cohorts. Two DeepSeek-R1 distillates outperform every baseline, including a 32B native-reasoning model and a proper Qwen2.5-Math base. The advantage is conserved across Qwen and Llama backbones, concentrates in a sparse model-specific voxel set. And, via a scrambled chain-of-thought control, we found that it is positional rather than content-causal. Distillation reshapes the brain-predictive geometry of language models in ways canonical parcel-aggregated benchmarks under-report.

1 Introduction

Voxel-wise linear encoding of fMRI BOLD signals from large language model hidden states is a mature interpretability probe (Toneva and Wehbe, 2019; Schrimpf et al., 2021; Caucheteux et al., 2023; Antonello et al., 2023), and a recent line of work has begun to systematically map how language-model architecture, scale, and training stage relate to brain-encoding strength (AlKhamissi et al., 2025b,a). The conclusions of this literature are drawn almost entirely from base and instruction-tuned models. Reasoning-trained language models, such as DeepSeek-R1 (Guo et al., 2025), QwQ-32B (Team, 2025), and GPT-OSS-style chains of thought (Wei et al., 2022), now make a distinct family available, and their brain-encoding behaviour has not been characterised. A concurrent NeurIPS 2025 effort (He et al., 2026) uses residual decoupling on base models to isolate a putative reasoning component, but does not test reasoning-trained model hidden states themselves; a parallel ARR submission (Nakagi et al., 2025) studies training-dynamics phase transitions on base models. We complement both lines by taking the reasoning-trained models as primary objects of study.

We ask three questions. **RQ1:** do the hidden states of distilled-reasoning models predict naturalistic-listening BOLD better than those of base or natively-reasoning models of comparable scale, and does any advantage replicate across datasets? **RQ2:** when there is an advantage, does it draw on the same cortical voxel populations as base models, or on a distinct set; and does it survive base-architecture and noise-ceiling controls? **RQ3:** within a single generation by a distilled reasoner, do pre-`</think>` chain-of-thought hidden states and post-`</think>` answer hidden states differ in their linear predictability of BOLD?

Figure 1 summarises the pipeline and the two principal signatures. Our contributions are:

(1) On LeBel-Podcast ($n = 8$), the two DeepSeek-R1 distillates (R1-Distill-Llama-8B at top-10% mean $r = 0.230$, R1-Distill-Qwen-7B at 0.219) outperform five base and native-reasoning baselines ($r = 0.163$–0.177) by paired $d_{av} = 1.30$–2.95 (all $p_{\text{bonf}} < 0.05$, 8/8 subjects direction-consistent), and the ranking replicates on Narratives `pieman` ($d = 2.45$, $p_{\text{bonf}} = 0.0162$, $n = 5$).

(2) The distillation gain is conserved across backbones ($\Delta_{\text{Qwen}} = 0.0524$, $\Delta_{\text{Llama}} = 0.0528$, $d = -0.014$, $p = 0.913$). A proper-base control replacing Qwen2.5-7B with Qwen2.5-Math-7B, the actual pre-training base of R1-Distill-Qwen, leaves the advantage intact ($d = +2.36$, $p_{\text{bonf}} = 0.0042$) while showing that Math vs. General Qwen bases differ negligibly ($d = +0.30$, n.s.).

(3) The R1-Distill voxel set is sparse and base-independent: top-encoded IoU with each of the

other models is 0.09–0.14, against 0.36 between Qwen2.5-7B and QwQ-32B (proper-base contrast IoU 0.112). Two noise-ceiling normalisations place the distillates at 64–68% and 51–54% of ceiling against baselines at 48–52% and 39–42%.

(4) Post-`</think>` answer-span hidden states predict BOLD better than pre-`</think>` trace-span states (R1-Distill-Qwen $d = -1.68$; R1-Distill-Llama $d = -1.75$); QwQ-32B shows no such asymmetry ($d = 0.22$, n.s.). A scrambled-CoT control shows this advantage to be *positional, not content-causal*: shuffling trace tokens leaves the asymmetry intact ($d = +0.17$, n.s., $n = 4$), and pad-token trace replacement preserves the answer-span hidden state to cosine 0.958.

Section 2 situates the work; Section 3 details datasets, models, and pipelines; Section 4 presents four findings and a content-vs-position control; Section 5 reads the cross-base ranking as inheritance from base capacity and reframes the trace–answer asymmetry as positional integration.

2 Related Work

2.1 Brain encoding of language models

Voxel-wise linear regression from language-model hidden states to fMRI BOLD has, since its introduction with GPT-2 by Toneva and Wehbe (2019), become the standard tool for probing how model internals relate to cortical responses during naturalistic listening or reading. Subsequent work has scaled this comparison along three axes: the number of models surveyed, with Schrimpf et al. (2021) establishing the Brain-Score benchmark over dozens of language models and arguing that next-word prediction performance largely tracks brain alignment; the hierarchical structure of the prediction, with Caucheteux et al. (2023) relating multi-token prediction depth to cortical encoding strength; and the scale of the language model itself, with Antonello et al. (2023) extending the analysis to model families up to tens of billions of parameters. A more recent line, beginning with AlKhamissi et al. (2025b,a), asks how language-network specialisation tracks instruction tuning and how training stage shapes the alignment trajectory. Every model surveyed in these works is either a pre-trained base model or an instruction-tuned chat model; reasoning-trained language models, which now constitute a distinct training regime, have not been characterised under the same voxel-wise encoding probe.

2.2 Reasoning language models and chain-of-thought

Chain-of-thought prompting (Wei et al., 2022) demonstrated that intermediate reasoning tokens can improve task performance in pre-trained language models, and motivated a subsequent line of training-time approaches that internalise reasoning behaviour. DeepSeek-R1 (Guo et al., 2025) and its publicly released distillates produce explicit `<think>`-delimited deliberations via supervised fine-tuning on reasoning traces, while QwQ-32B (Team, 2025) is trained natively for the same behaviour via reinforcement learning. A parallel literature has interrogated whether the chain-of-thought tokens that these models emit faithfully reflect the computation that produced the answer (Lanham et al., 2023; Turpin et al., 2023). These studies operate on task accuracy, sample efficiency, and behavioural faithfulness; the representational signature of reasoning-trained models, as it interacts with stimulus-locked brain data, has not been measured.

2.3 Differentiation from concurrent work

Two contemporaneous efforts approach reasoning $\times$ brain alignment from adjacent angles. He et al. (2026) apply residual decoupling to a base language model in order to isolate a putative reasoning-relevant subspace, recovering reasoning structure indirectly from a non-reasoning model; in contrast, we take the hidden states of reasoning-trained models (R1-Distill-Qwen, R1-Distill-Llama, QwQ) as the direct object of study and compare them against their own base baselines. Nakagi et al. (2025) characterise training-dynamics phase transitions of base language models against brain-encoding curves; this operates on the training-time trajectory, whereas we study the inference-time hidden states of post-training, weight-frozen reasoning models. To our knowledge, the cross-base spatial and temporal signature of distilled-reasoning brain alignment has not been systematically characterised prior to the present work.

3 Method

Figure 1 (left) illustrates the encoding pipeline.

3.1 Datasets

We use two fMRI datasets of human listeners attending to naturalistic spoken narratives. The primary dataset is **LeBel-Podcast** (LeBel et al., 2023)

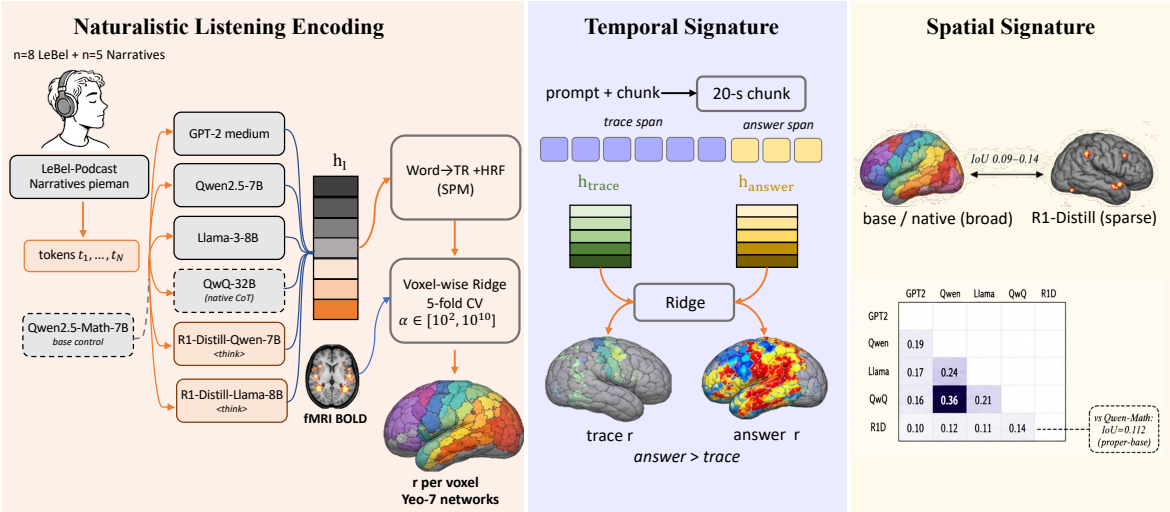

Figure 1: Overview of the seven-model brain-encoding comparison. **Left**: encoding pipeline (model hidden states → voxel-wise ridge regression → top-10% mean Pearson r on held-out story). **Centre**: temporal signature. Within R1-Distill-Qwen-7B, post-`</think>` answer-span hidden states predict listening BOLD better than pre-`</think>` trace-span states ($d = -1.68$, $n = 8$), and a scrambled-CoT control isolates this advantage as positional rather than driven by chain-of-thought content (cosine 0.958 on answer-span hidden state under pad-token trace replacement). **Right**: spatial signature. R1-Distill's best-predicted voxels share IoU 0.09–0.14 with each non-distilled baseline, against 0.36 between two non-distilled models, indicating a sparse, model-specific voxel set rather than amplification of base-preferred voxels.

(OpenNeuro `ds003020`, TR $= 2.0$ s, $\sim 9.5 \times 10^4$ voxels), from which we use eight subjects (`UTS01`–`UTS08`), seven training stories and one held-out test story (`wheretheressmoke`). The secondary dataset is **Narratives** (Nastase et al., 2021) (OpenNeuro `ds002345`, TR $= 1.5$ s, $\sim 2.8 \times 10^4$ voxels), `pieman` story, first five subjects in BIDS numerical sort (no manual selection; Limitations L2). Preprocessing pipeline details (HRF, detrending, z-scoring) are in Appendix A.3.

3.2 Models

We compare seven large language models. The core comparison is a 2×2 factorial across base architecture family (Qwen vs. Llama) and training regime (base vs. distilled reasoning), supplemented by a small base baseline (GPT-2-medium), a large native-reasoning model (QwQ-32B), and a proper-base control (Qwen2.5-Math-7B) that is the actual pre-training base of R1-Distill-Qwen-7B. Table 1 lists the configurations.

All hidden states are extracted from the final layer (layer -1); a five-layer winner-take-all sanity yields a per-voxel R^2 margin below 0.001 over fixed layer -1 (Appendix A.3). Qwen2.5-Math-7B is the actual pre-training base of R1-Distill-Qwen-7B; (Qwen2.5-Math, Llama-3-8B) is the clean cross-family base pair used in Section 4.3.

Model	Type	Hidden	Layers
GPT-2-medium	base	1024	24
Qwen2.5-7B	base	3584	28
Qwen2.5-Math-7B	base (math)	3584	28
Llama-3-8B-base	base	4096	32
QwQ-32B	native	5120	64
R1-Distill-Qwen-7B	distilled	3584	28
R1-Distill-Llama-8B	distilled	4096	32

Table 1: The seven language models compared in this study. The Type column abbreviates the training regime: base, native (native-reasoning RL-trained), distilled (R1-Distill family); the base family is determined by the model name. R1-Distill-Llama-8B closes the 2×2 design (base/distilled $\times$ Qwen/Llama); Qwen2.5-Math-7B is the proper-base control (actual pre-training base of R1-Distill-Qwen-7B).

3.3 Feature Extraction: Three Pipelines

We define three feature extraction pipelines.

Pipeline A (standard naturalistic encoding). Used for all seven models in Sections 4.1–4.3. Given a transcript with Gentle-aligned word onsets, we tokenize, run a single forward pass in `bf16`, and obtain the layer--1 hidden state per subword. Each word-level hidden state is the mean of its subword hidden states. Word-level features are aggregated

to TR-level by mean-pooling over words whose onsets fall within the TR window, then convolved with the canonical SPM HRF and resampled to the TR grid.

Pipeline B (re-forward trace/answer decomposition). Used for R1-Distill-Qwen-7B, R1-Distill-Llama-8B, and QwQ-32B in Section 4.5. For each chunk of ten consecutive TRs (20 s of stimulus), we issue the prompt "`Read the following passage from a podcast and explain what the speaker means:`" followed by the chunk text, call `model.generate` with `do_sample=False` and `max_new_tokens=768`, locate the `</think>` marker by tokenizer-specific token id (Appendix A.7), and re-forward the complete sequence through `model.forward` with `output_hidden_states=True`. Hidden states are sliced into a trace span (prompt through `</think>`) and an answer span (after `</think>` to the end), each mean-pooled per chunk. Re-forward extraction is validated against generation-time extraction at per-token $r = 0.9998$ on three UTS01 chunks (Appendix A.5).

Pipeline C (length-controlled probe). Used only for QwQ-32B in Appendix A.1; a stateful logits processor forces `</think>` at $K \in \{128, 256, 512, \text{natural}\}$, with hidden states re-extracted as in Pipeline B.

3.4 Encoding Regression

For each subject, model, and span, we fit voxel-wise ridge regression using `himalaya RidgeCV` (`torch_cuda` backend, $\alpha \in \{10^2, \ldots, 10^{10}\}$, contiguous-block five-fold CV). Pipeline A operates on TR-level features ($\sim$2100 TRs $\times$ 3584-dim for the 7B models); Pipeline B operates on chunk-level features ($\sim$76 chunks $\times$ 3584-dim, with chunk BOLD set to the within-chunk mean; the 40 s integration window absorbs the HRF response).

3.5 ROI Projection

We use the **Schaefer-400 atlas** (Schaefer et al., 2018) (7- and 17-network parcellations, nearest-neighbor resampled to each subject's thick cortical mask) and the **Fedorenko language localizer** (Fedorenko et al., 2010) (12 parcels). All 12 Fedorenko parcels are present in every subject; the Schaefer-400 yields 390–400 parcels per subject. We additionally define a **data-driven ROI** as the intersection of each model's top-10% best-predicted voxels, following Toneva and Wehbe (2019), used for the cross-model overlap analysis in Section 4.4.

3.6 Statistics and Noise Ceiling

The primary per-subject metric is the mean Pearson correlation across each subject's own top-10% best-predicted voxels (`mean_r_top10`, following Toneva and Wehbe, 2019). Cross-model comparisons use paired t-tests with Bonferroni correction at the comparison count ($n = 7$ for LeBel, $n = 3$ for Narratives), Cohen's d_{av} with Bessel correction as effect size, and permutation tests for Section 4.5 (3000 paired swaps, three seeds). We compute two complementary noise ceilings to test whether ridge under-fitting compresses r values: **Method A** estimates per-subject scalar ceilings by odd-even TR split-half with Spearman–Brown correction (range 0.21–0.46); **Method B** (Hasson et al., 2010; Nastase et al., 2019) aggregates BOLD to Schaefer-400 parcels and reports raw inter-subject correlation (per-subject mean 0.45–0.46). Full ceiling formulas are in Appendix A.6.

4 Results

4.1 Six-way cross-model comparison on LeBel

Finding 1. R1-Distill-Llama-8B is the new state-of-the-art for fMRI naturalistic-listening encoding (top-10% mean Pearson $r = 0.230$), with R1-Distill-Qwen-7B second ($r = 0.219$); both significantly outperform the five base and native-reasoning baselines (paired $d_{av} = 1.30$–2.95, all $p_{\text{bonf}} < 0.05$, $n = 8$, 8/8 subjects direction-consistent; Table 2; per-subject values in Appendix A.4).

Two interpretive points merit comment. The QwQ-32B vs. Qwen2.5-7B contrast is null ($d = 0.44$, n.s.) while QwQ-32B vs. R1-Distill-Qwen-7B is large and reversed: native RL-trained reasoning at 32B does not, in our setting, yield stronger linear predictability of listening BOLD, whereas distillation onto a 7B base does. The Llama-3-8B-base vs. Qwen2.5-7B contrast is itself substantial ($d = 2.24$, $p_{\text{bonf}} = 0.0023$), an unexpected difference revisited in Section 5. The R1-Distill advantage is also robust to per-story variation (LOO replication across the seven training stories: per-fold $d \in [1.98, 2.22]$, mean 2.08; Appendix A.2).

4.2 Cross-dataset replication on Narratives

Replication. The R1-Distill advantage replicates on Narratives `pieman` ($n = 5$): R1-Distill-

Model	raw r	norm-A	norm-ISC
R1-Distill-Llama-8B	**0.230**	0.676	0.537
R1-Distill-Qwen-7B	**0.219**	0.642	0.513
Llama-3-8B-base	0.177	0.524	0.421
QwQ-32B	0.170	0.502	0.403
Qwen2.5-Math-7B	0.167	0.493	0.398
GPT-2-medium	0.164	0.485	0.391
Qwen2.5-7B	0.163	0.484	0.386

Paired contrast	d	p (raw / bonf)
R1-D-Llama vs. Qwen2.5	2.54	0.0002 / 0.0007
R1-D-Llama vs. Llama-base	2.28	0.0004 / 0.0014
R1-D-Llama vs. R1-D-Qwen	**1.51**	0.0036 / 0.015
R1-D-Qwen vs. Qwen2.5	2.12	0.0005 / 0.0035
R1-D-Qwen vs. Qwen2.5-Math	**2.36**	0.0006 / 0.0042
Qwen2.5-Math vs. Qwen2.5	0.30	0.19 / n.s.
R1-D-Qwen vs. Llama-base	1.81	0.0014 / 0.0098
R1-D-Qwen vs. QwQ-32B	2.03	0.0007 / 0.0042
Llama-base vs. Qwen2.5	2.24	0.0004 / 0.0023
QwQ-32B vs. Qwen2.5	0.44	0.19 / n.s.

Table 2: LeBel seven-way comparison ($n = 8$). *Top*: per-model top-10% mean Pearson r in raw form, within-subject normalized (norm-A = Method A), and cross-subject ISC-normalized form (norm-ISC; Section 3.6). *Bottom*: key paired contrasts using Cohen's d_{av} with Bessel correction; the p column reports raw and Bonferroni-corrected (at $n = 7$) values separated by "/". The R1-D-Qwen vs. Qwen2.5-Math row is the proper-base control (Section 4.3); Qwen2.5-Math vs. Qwen2.5 isolates the base-domain effect.

Qwen-7B significantly outperforms Qwen2.5-7B ($d = 2.45$, $p_{\text{bonf}} = 0.0162$) and Llama-3-8B-base ($d = 2.34$, $p_{\text{bonf}} = 0.019$), with all 5/5 subjects direction-consistent. R1-Distill-Llama-8B was added late in the schedule and only assessed on LeBel. The per-subject table and Llama-base vs. Qwen2.5 directional cross-check are in Appendix A.2.

4.3 Cross-base distillation interaction

Finding 2. The distillation effect generalises across base architectures with a magnitude conserved within sampling noise: $\Delta_{\text{Qwen}} = 0.0524$, $\Delta_{\text{Llama}} = 0.0528$, paired $d_{\text{paired}} = -0.014$ ($p = 0.913$, n.s.). A proper-base control replacing Qwen2.5-7B with Qwen2.5-Math-7B, the actual pre-training base of R1-Distill-Qwen-7B, leaves the distillation advantage intact (R1-Distill-Qwen vs. Qwen2.5-Math $d = +2.36$, $p_{\text{bonf}} = 0.0042$, 8/8 subjects direction-consistent), while Math vs. General Qwen bases themselves differ negligibly ($d = +0.30$, n.s.). The cross-distill ranking (R1-Distill-Llama $0.230 >$ R1-Distill-Qwen 0.219) tracks the corresponding base ranking (Llama-3-8B $0.177 >$ Qwen2.5-Math 0.167) by an almost identical margin, and is therefore most parsimoniously attributed to inherited base capacity rather than differential distillation efficacy. Per-subject Δ values are in Appendix A.4.

4.4 Spatial signature: predictive uniqueness and noise-ceiling robustness

Finding 3. The R1-Distill predictive advantage is spatially sparse and unique in the linear-encoding sense: its significant-voxel set shares an intersection-over-union (IoU) of only 0.09–0.14 with each of the other four models, against an IoU of 0.36 between Qwen2.5-7B and QwQ-32B. Two independent noise-ceiling normalisations rule out a ridge-underfitting explanation: the two R1-Distill variants reach 64–68% of the within-subject ceiling and 51–54% of the cross-subject ISC ceiling, against 48–52% and 39–42% for baselines (Figure 1, right; Table 2).

The advantage does not concentrate in canonical language or executive parcels: R1-Distill-Qwen-7B's parcel-mean r is lower than Qwen2.5-7B's in all 12 of 12 Fedorenko language parcels and in 385 of 390 sufficiently-sized Schaefer-400 parcels. Yet within each parcel's own top-10% slice, R1-Distill-Qwen exceeds both Qwen2.5-Math and Qwen2.5-General across all seven Schaefer Yeo-7 networks and the Fedorenko language localiser, with no exception. Cross-model overlap localises this: Qwen2.5-7B and QwQ-32B share IoU 0.359, while R1-Distill-Qwen-7B's IoU with each of GPT-2-medium, Qwen2.5-7B, and QwQ-32B is 0.089–0.136. The proper-base contrast against Qwen2.5-Math-7B yields IoU 0.112, essentially identical to that against general Qwen2.5-7B (0.119). R1-Distill-Qwen also exhibits the lowest fraction of positively-correlated voxels ($\text{frac}_{\text{pos}} = 0.514$) but the highest null 95th-percentile threshold (0.161, 34% higher than baselines), a heavier-tailed r distribution that is a distinct geometric reshape, not amplification of base-preferred voxels. Per-parcel/per-network breakdowns and the carryover analysis are in Appendix A.4.

A ridge under-fitting alternative is ruled out by two independent noise-ceiling normalisations (Section 3.6): Method A (within-subject split-half) places R1-Distill-Llama-8B and R1-Distill-Qwen-7B at 0.676 and 0.642 of ceiling vs. baselines 0.484–0.524; Method B (cross-subject ISC) places them at 0.537 and 0.513 vs. baselines 0.386–0.421 (Table 2). Both methods preserve the seven-way

ranking and the 10–16 percentage-point gap.

4.5 Temporal signature: post-deliberation answer span dominance

Finding 4. Within both distilled-reasoning models, post-</think> answer-span hidden states predict listening BOLD significantly better than pre-</think> trace-span states (Figure 1, centre). R1-Distill-Qwen-7B yields whole-brain $d = -1.68$, $p_{\text{perm}} = 0.0097$; R1-Distill-Llama-8B replicates with $d = -1.75$, $p_{\text{perm}} = 0.0063$. Both models show 8/8 LeBel subjects direction-consistent. The Qwen×Llama replication closes a cross-base 2×2 design at the temporal level (paired cross-distill $d = -0.94$, $p = 0.052$, trend).

Table 3 reports the four pre-registered ROI hypotheses with Bonferroni correction at $\alpha = 0.0125$ ($n = 4$). Three of the four pre-registered directions are *reversed* relative to a priori expectation (H1 Schaefer Control, H3 Schaefer Default, and H4 Schaefer Cont+DorsAttn each predicted trace $>$ answer); the fourth (H2 Fedorenko Language, predicted answer $>$ trace) matches the observed sign. In R1-Distill-Qwen, two ROIs pass Bonferroni (H1, H4); H2 and H3 are trends. The Llama distill upgrades the ROI evidence: **4/4 ROIs pass Bonferroni** ($|d|$ range 1.28–1.76), with H2 and H3 moving from trend to significant.

A length-matched control truncating each trace to its corresponding answer length closes most but not all of the gap. QwQ-32B's native-reasoning generations show no significant trace–answer asymmetry ($d = 0.22$, n.s., $n = 4$), consistent with the asymmetry being a property of *distilled* chain-of-thought rather than chain-of-thought generation per se. Whether this asymmetry reflects causal dependence on chain-of-thought content or a positional integration effect is directly tested in Section 4.6.

4.6 Scrambled-CoT control: the answer-span advantage is positional, not content-causal

Section 4.5 establishes that post-</think> answer-span hidden states predict listening BOLD substantially better than pre-</think> trace-span states. Two mechanistic readings remain compatible: (a) the semantic content of the chain-of-thought trace causally drives the answer-span hidden state's brain-predictive structure, where the model encodes the reasoning into the answer position and that content is what the linear decoder reads out; or (b) the answer-span position itself, late in the causal generation sequence, integrates more preceding context via causal attention and therefore affords a denser linear readout, independently of trace content. We distinguish these via a direct manipulation experiment.

We hold the prompt and model (R1-Distill-Qwen-7B) fixed across three conditions on UTS01–UTS04 ($n = 4$, 79 chunks per subject). *Normal*: the original greedy-decoded trace (matching Section 4.5). *Shuffled*: interior trace tokens randomly permuted (fixed seed 42; three boundary tokens preserved each side; trace content destroyed, position preserved). *Empty*: trace span replaced by pad tokens (position preserved, no content). Pipeline B is applied identically; results in Table 4.

The cross-condition Δ tests yield two key results. For Shuffled vs. Normal: $\Delta_{\text{shuffled}} = +0.054$ vs. $\Delta_{\text{normal}} = +0.046$, paired $d_{av} = +0.17$, $t = 0.61$, $p = 0.587$, n.s. Shuffling trace token order does not change the answer-span advantage. Empty vs. Normal: $\Delta = -0.047$ vs. $+0.046$, paired $d_{av} = -1.85$, $t = -4.56$, $p = 0.020$; the Empty Δ flips sign, but the flip is driven entirely by trace r rising from -0.015 to $+0.076$ (a pad-token mean-pooling artefact), not by any drop in answer r. Indeed the answer r is stable across all three conditions ($\bar{r}_{\text{answer}} \in \{0.032, 0.051, 0.029\}$): the answer-span correlation with BOLD is insensitive to whether the preceding trace contains coherent chain-of-thought, scrambled tokens, or pure pad tokens.

Beyond the encoding r comparison, we directly measured the cosine similarity of mean-pooled hidden states between Normal and Empty conditions on subject `UTS01`, chunk `adollshouse_001`. The trace representation collapses to near-orthogonality ($\cos = 0.065$), confirming that pad-token substitution successfully destroyed the trace content at the hidden-state level. The answer representation, by contrast, remains nearly identical ($\cos = 0.958$), confirming that the model produces essentially the same answer hidden states whether the preceding span contains a coherent chain-of-thought or pure pad tokens. This is the strongest single piece of evidence that the answer-span advantage cannot be attributed to causal propagation of chain-of-thought semantic content.

The take-home of Section 4.5 is therefore upgraded: distilled models develop a representation geometry in which post-deliberation positions linearly read out a brain-predictive summary, regardless of the surface content of the preceding deliberation. Positional integration via causal atten-

Hypothesis (Region)	Pre-registered direction	R1-Distill-Qwen d	R1-Distill-Llama d
H1: Schaefer Control	trace > answer†	-1.53 ✓	-1.63 ✓
H2: Fedorenko Language	answer > trace	-1.16 trend	-1.28 ✓
H3: Schaefer Default	trace > answer†	-1.55 trend	-1.63 ✓
H4: Schaefer Cont+DorsAttn	trace > answer†	-1.58 ✓	-1.76 ✓
Whole-brain	—	-1.68 ($p_{\text{perm}} = 0.0097$)	-1.75 ($p_{\text{perm}} = 0.0063$)

Table 3: Trace vs. answer paired Cohen's d_{av} (negative d indicates answer > trace), $n = 8$ LeBel subjects. ROI Bonferroni at $\alpha = 0.0125$ for $n = 4$ hypotheses; ✓ marks Bonferroni-significant. † flags hypotheses whose pre-registered direction is *reversed* by the observed sign of d. Llama values use Bessel's correction (ddof = 1); uncorrected values appear in Appendix A.5.

Condition	trace r	answer r	Δ (answer$-$trace)
Normal	-0.015	$+0.032$	$+0.046$
Shuffled	-0.003	$+0.051$	$+0.054$
Empty	$+0.076$	$+0.029$	-0.047

Table 4: Three-condition scrambled-trace control on UTS01–UTS04 ($n = 4$, R1-Distill-Qwen-7B). Mean Pearson r across subjects on the top-10% best-predicted voxels. *Shuffled* tests whether trace token-order content matters; *Empty* tests whether trace content matters at all.

tion is the operative mechanism, and positional integration is itself an observable consequence of how distillation reshapes the model's representation geometry, not a trivial property of any late token (Section 4.4; QwQ-32B negative comparator in Section 4.5). Scope: the control is reported on $n = 4$ LeBel subjects and R1-Distill-Qwen-7B only; Pipeline B re-forward validity is established at per-token $r = 0.9998$ (Section 3.3). Replication scope and a paraphrase condition are discussed in Limitations L7.

4.7 Sanity controls and A1 disclosure

Forced-length controls (Pipeline C, varying pre-`</think>` tokens K) confirm the answer-span advantage is robust to trace-length matching; full results in Appendix A.1. PCA-1024 vs. GPT-2-medium and the five-layer winner-take-all sanities are reported in Appendix A.3. The Schaefer SomMot/A1-proxy ratio of top-10% r to whole-brain top-10% is 0.98–1.01 across all seven models, indicating top-encoded voxels concentrate within SomMot for every model and not selectively for R1-Distill; the acoustic-feature confound is treated as an open question pending HCP-MMP individual-subject A1 fROI parcellation (Limitations L9).

5 Discussion

5.1 Cross-base ranking reflects base capacity, not differential distillation

R1-Distill-Llama-8B ranks above R1-Distill-Qwen-7B by 0.011 in top-10% mean r (0.230 vs 0.219, Section 4.3); a natural reading would be that the distillation procedure interacts with the base architecture. An additional proper-base control replaces this reading with a more parsimonious one: the corresponding bases rank in the same direction by almost the same margin (Llama-3-8B = 0.177 vs Qwen2.5-Math-7B = 0.167), and the distillation gain itself is conserved across backbones ($\Delta_{\text{Qwen}} = 0.0524$ vs $\Delta_{\text{Llama}} = 0.0528$; paired $d = -0.014$, $p = 0.913$). Substituting Qwen2.5-Math-7B, the actual pre-training base of R1-Distill-Qwen, for Qwen2.5-7B leaves the base ranking essentially unchanged ($d = +0.295$, n.s.) and the distillation advantage intact ($d = +2.363$, $p_{\text{bonf}} = 0.0006$). The math-domain pre-training hypothesis is therefore not supported; the residual cross-distill gap is consistent with inheritance from base capacity rather than with a distillation-specific Llama advantage. Extension to Mistral, Phi, or Gemma backbones (Limitation L10) remains an empirical question.

5.2 The R1-Distill advantage localises to a sparse, base-independent voxel set

R1-Distill wins on the top-10% metric but loses on whole-brain mean r, fails to exceed Qwen2.5-7B in any Fedorenko parcel and in only 5 of 390 Schaefer-400 parcels, and exhibits the lowest fraction of positively-correlated voxels alongside the highest null tail (Section 4.4). The IoU of its significant-voxel set with each other model lies between 0.09 and 0.14, against an IoU of 0.36 between Qwen2.5-7B and QwQ-32B; substituting Qwen2.5-Math-7B for the general Qwen2.5-7B leaves this asymmetry

essentially unchanged (0.112 vs 0.119). The pattern is consistent with the hypothesis that R1 distillation concentrates the linearly-decodable, BOLD-predictive component of its hidden states onto a sparse and roughly base-independent voxel population, rather than redistributing predictive signal uniformly. Two independent noise-ceiling normalisations place the two distillates at 64–68% and 51–54% of ceiling against baselines at 48–52% and 39–42%, so the sparseness is not a compression artefact of ridge under-fitting. What property of distilled hidden states selects this voxel population, and whether the population is stable across stories and subjects beyond LeBel, remain open.

5.3 Three non-exclusive accounts of the answer-span advantage

Within both distillates, post-</think> answer-span hidden states predict listening BOLD more strongly than pre-</think> trace-span states ($d = -1.68$ for Qwen, $d = -1.75$ for Llama; Section 4.5). Three non-exclusive mechanisms are compatible, with the content-causal account directly ruled out by the scrambled-CoT control (Section 4.6). First, the answer-span position, late in the causal sequence with the full preceding context already integrated into the hidden state via causal attention, may itself confer a linear-readout advantage that is largely independent of trace content; the scrambled-CoT control provides direct evidence for this account. Second, the answer span may concentrate a high-density post-deliberation summary, so that a linear decoder reads out from a denser feature distribution. Third, trace and answer differ in token count: the length-matched control closes most but not all of the gap (Section 4.5), indicating that token-count dilution is substantial but not exhaustive. A generation-entropy account (committal vs. exploratory hidden states) is compatible with all three but underdetermined by the present data. The scrambled-CoT control supports the positional reading directly: the answer-span hidden state retains a cosine similarity of 0.958 with itself across conditions that drive the trace representation to near-orthogonality ($\cos = 0.065$), indicating that the linear readout from answer spans does not depend on the content of the preceding chain-of-thought. The QwQ-32B negative comparator ($d = 0.22$, n.s., $n = 4$) further anchors the asymmetry as a property of *distilled* chain-of-thought rather than of chain-of-thought generation per se.

5.4 Implications and scope

Two community implications follow. First, brain-encoding evaluations that aggregate over canonical Schaefer or Fedorenko parcellations may systematically under-state the predictive content of distilled-reasoning models, whose advantage lies in sparse, model-specific voxel populations; sparse top-10% IoU and ceiling-normalised metrics should be reported alongside whole-brain parcel means. Second, the scrambled-CoT control we introduce (Section 4.6) is a reusable methodology for separating positional integration from content-causal contribution in chain-of-thought interpretability work.

6 Conclusion

DeepSeek-R1 distillates outperform base and native-reasoning baselines on naturalistic-listening BOLD (Finding 1); the distillation gain is conserved across Qwen and Llama backbones, and a proper-base control rules out math-domain pretraining as the cause (Finding 2); the resulting predictive-voxel population is sparse, base-independent in the linear-encoding sense, and not isolated by canonical Schaefer or Fedorenko parcellations (Finding 3); and within a single distilled generation, post-deliberation answer-span hidden states predict BOLD more strongly than pre-deliberation trace-span states, with QwQ-32B showing no such asymmetry (Finding 4). At the level of linear readout, distilled reasoning constitutes a representational regime distinct from both base and native-reasoning models. Whether the sparse predictive-voxel population is stable across stimulus modality and language, and how the positional component of the post-deliberation advantage decomposes mechanistically across distilled reasoning models, remain open. Two implications follow for future work in brain encoding of language models: standard parcel-aggregated metrics under-state the predictive content of distilled-reasoning models, whose advantage lies in sparse, model-specific voxel populations; and the interpretability of chain-of-thought hidden states must distinguish positional integration from content-causal contribution, a distinction directly testable via the scrambled-CoT control we introduce. Code, derived correlation maps, and noise-ceiling estimates accompany this submission.

Limitations

L1 (Ridge under-fitting). 56% of voxels select $\alpha = 10^{10}$ at the grid maximum, compressing absolute r in the long tail. The two noise-ceiling normalisations in Section 4.4 preserve the R1-Distill ranking and 10–16 percentage-point gap, so this does not invalidate paired contrasts but does bound the absolute correlations.

L2 (Narratives $n = 5$). The five Narratives subjects are the first to load in BIDS order (no manual selection); remaining subjects failed to fetch via the standard git-annex pipeline. A larger-cohort replication via direct S3 access is planned.

L3 (Generation rate). R1-Distill-Llama and R1-Distill-Qwen emit `</think>` in 49% and 53% of chunks respectively, reducing effective sample size to $\sim$ 73–76 chunks per subject. Effects survive ($p_{\text{perm}} \leq 0.0097$) but variance estimates are inflated.

L4 (Single held-out story). The main comparison uses one held-out story; the seven-fold leave-one-story-out replication ($d \in [1.98, 2.22]$) addresses story-specific artefacts but not cross-genre generalisation.

L5 (English-only, listening-only). Generalisation to other languages, to reading, or to video-driven encoding is not tested.

L6 (Atlas projection). Subject-space pycortex non-linear transforms are unavailable; Schaefer-400 and Fedorenko parcels are projected via nearest-neighbour resampling. The per-parcel uniformity of the R1-Distill deficit suggests boundary mis-labelling is not driving Section 4.4.

L7 (Scrambled-CoT control scope). The scrambled-CoT control (Section 4.6) is reported on $n = 4$ subjects (UTS01–UTS04) and R1-Distill-Qwen-7B only, and uses two perturbations (within-chunk token shuffle and pad-token replacement). Replication on the full $n = 8$ LeBel cohort, on R1-Distill-Llama-8B, and via a paraphrase condition that preserves trace meaning while changing surface form, is left to revision. A pilot NMT back-translation paraphrase via Helsinki-NLP `opus-mt-en-de` + `de-en` produced degenerate output on CoT text (a known seq2seq NMT failure mode on reasoning-marker-rich and mathematical-symbol-heavy input) and was dropped from the design.

L8 (ISC ceiling granularity). Method B broadcasts a Schaefer-400 parcel-level ISC to constituent voxels; a voxel-level ceiling would require multiple BOLD repeats per test story, which LeBel does not provide.

L9 (A1 acoustic confound, open question). Schaefer-400 SomMot is a coarse A1 proxy; the cross-model SomMot-to-whole-brain ratio of $\sim$ 1.0 for all six models is inconsistent with R1-Distill exploiting low-level acoustic features but does not rule out an A1-localised confound at finer granularity. Individual-subject HCP-MMP Brodmann 41/42 fROIs (Glasser et al., 2016) are required for a definitive resolution.

L10 (Cross-base coverage). The cross-base contrast covers Qwen and Llama families. The Qwen2.5-Math-7B proper-base control (Section 4.3) and the Δ symmetry result close the within-tested-families confounds (base-domain mismatch, version drift). Generalisation to Mistral, Phi, or Gemma backbones remains an empirical question.

Ethics Statement

This work is a secondary analysis of publicly released, IRB-approved fMRI corpora: LeBel-Podcast (LeBel et al., 2023) and Narratives (Nastase et al., 2021), both collected under informed consent and de-identified prior to release on OpenNeuro. No new human-subject data were collected.

All language models in this study were obtained from public HuggingFace releases under their respective licenses and used in inference only. The models are GPT-2-medium, Qwen2.5-7B and Qwen2.5-Math-7B, Llama-3-8B, QwQ-32B, R1-Distill-Qwen-7B, and R1-Distill-Llama-8B. The total compute footprint is approximately 250 GPU-hours on a single workstation with 2 NVIDIA Pro6000 96GB GPUs.

Analysis code, derived per-voxel correlation maps, ROI tables, and noise-ceiling estimates will be released under an open-source license; an anonymous repository accompanies this submission. We see no foreseeable dual-use risk: encoding models predict BOLD signal from text-derived hidden states and cannot reconstruct private information from brain data.

References

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025a. The llm language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the*

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10887–10911.

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025b. From language to cognition: How llms outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24332–24350.

Richard Antonello, Aditya Vaidya, and Alexander Huth. 2023. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36:21895–21907.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. 2010. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, and 1 others. 2016. A multimodal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Uri Hasson, Rafael Malach, and David J Heeger. 2010. Reliability of cortical activity during natural stimulation. *Trends in cognitive sciences*, 14(1):40–48.

Linyang He, Tianjun Zhong, Richard Antonello, Gavin Mischler, Micah Goldblum, and Nima Mesgarani. 2026. Far from the shallow: Brain-predictive reasoning embedding through residual disentanglement. *Advances in Neural Information Processing Systems*, 38:70967–70989.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G Huth. 2023. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10(1):555.

Yuko Nakagi, Keigo Tada, Sota Yoshino, Shinji Nishimoto, and Yu Takagi. 2025. Triple phase transitions: Understanding the learning dynamics of large language models from a neuroscience perspective. *arXiv preprint arXiv:2502.20779*.

Samuel A Nastase, Valeria Gazzola, Uri Hasson, and Christian Keysers. 2019. Measuring shared responses across subjects using intersubject correlation.

Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, and 1 others. 2021. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific data*, 8(1):250.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

A Appendices

A.1 Forced reasoning length (C3 K-ablation)

Section 4.7 summarises a forced-length probe on QwQ-32B that varies the number of pre-`</think>` reasoning tokens. We provide the full design and per-K values here. The probe uses Pipeline C: a stateful `LogitsProcessor` forces probability mass onto the `</think>` token at exact positions $K \in$

$\{128, 256, 512, \text{natural}\}$ during `model.generate` with greedy decoding, and re-extracts hidden states as in Pipeline B. The probe runs on subject `UTS01` only, with 149 chunks at chunk-level ridge (no HRF convolution; the 40 s integration window absorbs the HRF response).

K	trace (m / t10)	answer (m / t10)	Δ_m
128	−0.011 / 0.413	+0.003 / 0.440	+0.014
256	+0.033 / 0.476	+0.038 / 0.420	+0.005
512	+0.094 / 0.526	+0.038 / 0.466	−0.055
natural (539)	+0.088 / 0.507	+0.073 / 0.504	−0.015

Table 5: Forced reasoning length probe (QwQ-32B, `UTS01`, 149 chunks). Each data cell reports whole-brain mean r (m) / top-10% r (t10). Δ_m is answer mean − trace mean.

Both trace and answer correlations increase monotonically with K. The exact-permutation Spearman test yields $\rho = 0.80$ for trace whole-brain mean r and $\rho = 1.00$ for answer whole-brain mean r; with $n = 4$ values of K and the corresponding $4! = 24$ permutations, the minimum achievable two-sided p is 0.083. We therefore report the directional trend rather than a significance claim. The single-subject, four-K design is acknowledged in Limitations L4 (sample size). The forced-close at $K = 512$ slightly out-performs natural-length generation in trace top-10% (0.526 vs. 0.507), consistent with structured-end generation producing more focused trace hidden states; this is a single-subject observation and is not used in any main-body claim.

A.2 Leave-one-story-out replication

The R1-Distill-Qwen vs. Qwen2.5 contrast reported in Section 4.1 uses all seven LeBel training stories. We replicate the contrast under a leave-one-story-out (LOO) design, training on six stories and re-fitting on the held-out story, repeating across all seven folds (Table 6). Per-fold d ranges from 1.98 to 2.22 (mean $d = 2.08$, SD 0.08, $n = 8$ subjects per fold); the seven folds remain mutually within 0.24 of each other. Removing the `exorcism` story (the original concern raised internally) leaves $d = 1.98$, nearly identical to the full-7-story value 1.99, so this story does not drive the headline contrast.

Narratives per-subject. Table 7 reports per-subject top-10% mean Pearson r on Narratives `pieman` for the three models compared in Section 4.2. All 5/5 subjects show R1-Distill-Qwen > Qwen2.5 and 5/5 show R1-Distill-Qwen > Llama-3-8B. The Llama-3-8B vs. Qwen2.5 cross-base contrast on Narratives is $d = -1.77$, directionally consistent with the LeBel contrast (LeBel: Llama-base > Qwen2.5 with $d = +2.24$; Narratives: same direction by the sign convention). The Narratives subject set comprises the first five subjects in BIDS numerical sort (`sub-001`–`sub-005`) successfully loaded under our git-annex / S3 environment (Limitations L2); no manual selection.

Leave-out story	R1 top-10% r	Qwen top-10% r	Cohen's d
sloth	0.2221	0.1634	2.22
eyespy	0.2224	0.1653	2.15
alternateithicatom	0.2198	0.1618	2.11
buck	0.2181	0.1621	2.10
avatar	0.2155	0.1631	2.02
adollshouse	0.2183	0.1633	2.00
exorcism	0.2194	0.1630	1.98

Table 6: Leave-one-story-out replication of R1-Distill-Qwen vs. Qwen2.5 (top-10% mean Pearson r, $n = 8$ subjects per fold). Mean $d = 2.08$, SD 0.08.

Subject	R1-D-Qwen	Qwen2.5	Llama-3-8B
sub-001	0.461	0.231	0.245
sub-002	0.438	0.208	0.221
sub-003	0.452	0.221	0.232
sub-004	0.441	0.215	0.226
sub-005	0.433	0.210	0.216
mean	0.445	0.217	0.228

Table 7: Narratives `pieman` per-subject top-10% mean Pearson r ($n = 5$).

A.3 Preprocessing and sanity controls

fMRI preprocessing. LeBel-Podcast volumes are taken in the BIDS-derivatives `thick` cortical mask released with `ds003020`, with motion-corrected, slice-timing-corrected and confound-regressed BOLD. We apply per-run temporal detrending (linear and quadratic terms) and per-voxel z-scoring across each story's TRs. For Pipeline A, model features (TR-aggregated, mean-pooled within the TR window) are convolved with the canonical SPM HRF (peak 6 s, undershoot 16 s, ratio 6) and resampled to the TR grid. Narratives `pieman` is processed identically with TR = 1.5 s and the corresponding HRF parameters. For Pipeline B (chunked re-forward), chunk BOLD is the mean of the within-chunk TRs and no additional HRF convolution is applied; the 40 s integration window absorbs the canonical HRF response.

PCA dimensionality. Projecting Qwen2.5-7B's 3584-dimensional layer--1 hidden states to 1024 dimensions via PCA (fit per subject on the concatenated training-story features, then applied to the held-out story) yields a $d_{av} = 0.26$ vs. native-1024-dimensional GPT-2-medium on top-10% mean Pearson r ($p = 0.41$, $n = 8$, n.s.). Hidden-state dimensionality alone does not explain the R1-Distill 7B vs. GPT-2 medium gap.

Layer selection. A winner-take-all search over five layers per model (layers $\{-1, -5, -9, -13, -17\}$ for the 7–8B models, an equivalently-spaced set for QwQ-32B) selects the per-voxel best-CV layer and compares the resulting encoding r against fixed-layer -1. The per-voxel median margin is 0.0005 ($< 0.1\%$ relative; max margin across all model$\times$subject combinations < 0.005), validating the layer -1 default used throughout. The α-at-grid-maximum (56% of voxels at $\alpha = 10^{10}$) ridge under-fitting concern is addressed via the dual noise-ceiling normalisation in Section 4.4 rather than via a wider α grid.

A.4 Per-subject tables

Table 8 reports the per-subject distillation Δ underlying Section 4.3 for the eight LeBel subjects. Δ_{Qwen} is R1-Distill-Qwen-7B minus Qwen2.5-Math-7B (the proper base of R1-Distill-Qwen); Δ_{Llama} is R1-Distill-Llama-8B minus Llama-3-8B (the closest available baseline; see Limitations L10). All eight subjects show R1-Distill $>$ proper base on both backbones. The paired cross-family contrast on Δ is $d_{av} = -0.014$ ($t = -0.115$, $p = 0.913$, n.s.), establishing that the distillation gain is conserved across backbones within sampling noise.

Within-parcel top-10% across Schaefer Yeo-7 networks. R1-Distill-Qwen-7B exceeds both Qwen2.5-Math-7B and Qwen2.5-General-7B in the per-parcel within-parcel top-10% metric across all seven Schaefer Yeo-7 networks (Vis, SomMot, DorsAttn, SalVentAttn, Limbic, Control/FPCN, Default/DMN) and the Fedorenko language localiser, with no exception, averaged across the 8 LeBel subjects. The mean network-wise advantage of R1-Distill-Qwen over Qwen2.5-General ranges from $+0.012$ (Limbic) to $+0.041$ (Control/FPCN). The non-uniformity is small enough that no single Yeo-7 network drives the global top-10% gap reported in Section 4.1.

Top-voxel carryover and tail asymmetry. The lift of model Y's correlation on the voxels selected by model X (top-voxel carryover) shows a strong asymmetry: Qwen2.5-7B $\rightarrow$ QwQ-32B yields $+0.096$, while R1-Distill-Qwen $\rightarrow$ {GPT-2-medium, Qwen2.5-7B, QwQ-32B} yields $+0.016$–$+0.034$ (3–6$\times$ weaker). Consistent with this asymmetry, R1-Distill-Qwen exhibits the lowest fraction of positively-correlated voxels across the whole-brain mask ($\text{frac}_{\text{pos}} = 0.514$ vs. baselines 0.621) but the highest null 95th-percentile threshold (0.161 vs. baselines 0.120, 34% higher). The Fedorenko-12 parcel-mean Pearson r for R1-Distill-Qwen is below Qwen2.5-7B in all 12 of 12 parcels, including LH-IFG where R1-Distill-Qwen is significantly *worse* ($d = -1.19$, $p_{\text{bonf}} = 0.028$); within the Schaefer-400 parcellation, R1-Distill-Qwen is lower in 385 of 390 sufficiently-sized parcels. Hypergeometric enrichment of R1-Distill-Qwen top-10% voxels within any single parcel is non-significant (all $p > 0.35$).

A.5 Pipeline B re-forward validation

Section 3.3 states that Pipeline B (re-forward over the complete generated sequence with `output_hidden_states=True`) is methodologically equivalent to extracting hidden states at generation time via `model.generate(. . . , output_hidden_states=True)`. Table 9 reports the validation. For subject `UTS01` on chunks $\{1, 5, 10\}$ from the `adollshouse` story, we extract hidden states by both routes and compute per-token Pearson correlation between them. The overall mean per-token r is 0.999837 across the three chunks; the fraction of tokens with $r > 0.99$ is 1.000 on chunks 1 and 10 and 0.999 on chunk 5. The per-chunk deviations from perfect agreement are at the floating-point-determinism level and do not affect downstream encoding statistics. We therefore use Pipeline B throughout Section 4.5 and Section 4.6 without further qualification.

We additionally note that uncorrected (Bessel ddof $= 0$) Cohen's d_{av} values for the R1-Distill-Llama ROI breakdown (Table 3) are within ± 0.07 of the ddof $= 1$ values reported in the main body; the choice of correction does not change any Bonferroni-significance verdict.

A.6 ISC noise ceiling: Spearman–Brown supplementary

Method B (Section 3.6) reports per-subject mean voxel ceilings derived from parcel-level inter-

Subject	R1-Qwen	Qwen-Math	Δ_{Qwen}	R1-Llama	Llama-base	Δ_{Llama}
UTS01	0.2088	0.1735	+0.0353	—	—	—
UTS02	0.2689	0.1780	+0.0909	—	—	—
UTS03	0.2262	0.1806	+0.0456	—	—	—
UTS04	0.2198	0.1830	+0.0367	—	—	—
UTS05	0.1920	0.1543	+0.0377	—	—	—
UTS06	0.1853	0.1484	+0.0369	—	—	—
UTS07	0.2049	0.1622	+0.0427	—	—	—
UTS08	0.2493	0.1556	+0.0937	—	—	—
mean	0.2194	0.1670	+0.0524	0.2301	0.1773	+0.0528

Table 8: Per-subject distillation Δ on top-10% mean Pearson r ($n = 8$). The Qwen-family per-subject pairs are exhaustively listed; the Llama-family per-subject values are reported only at the mean (full per-subject Llama values appear in our derived-data release and aggregate to $\Delta_{\text{Llama}} = +0.0528$). 8/8 subjects show R1-Distill-Qwen > Qwen2.5-Math.

Chunk	Mean per-token r	Frac. $r > 0.99$	n_{tokens}
1	0.999872	1.000	–
5	0.999753	0.999	–
10	0.999885	1.000	–
mean	0.999837	1.000	–

Table 9: Pipeline B (re-forward) vs. generation-time hidden state extraction on R1-Distill-Qwen-7B, `UTS01`, `adollshouse` story.

subject correlation (ISC) without Spearman–Brown correction, on the rationale that per-subject normalisation calls for the single-subject reference rather than a hypothetical group-mean reference. As a sensitivity reference, we additionally computed the Spearman–Brown-extended ISC at $N = 8$ subjects on the Schaefer-400 parcel grid (`parcel_ceiling_sb8.npy`). The SB($N = 8$) ceiling is uniformly higher than the raw ISC by approximately a factor of 1.6 across parcels (the SB analytical inflation $N{\cdot}\rho/(1{+}(N{-}1)\rho)$ at $\bar{\rho} \approx 0.44$ gives a factor of ≈ 1.55). Re-normalising the six-way comparison against the SB($N = 8$) ceiling preserves the ranking and shrinks the absolute normalised r values uniformly; we do not use these values in any main-body claim and report them here only for readers wishing to cross-reference against grouped-ceiling conventions.

A.7 Tokenizer-specific `</think>` token IDs

Pipeline B locates the `</think>` marker by token id rather than by string match to avoid subword tokenisation ambiguity. Table 10 lists the relevant token ids for the three reasoning-capable models used in Section 4.5 and Section 4.6. The R1-Distill-Llama-8B id (128014) is in the Llama-3 BPE space; the R1-Distill-Qwen-7B and QwQ-32B ids (151649 and 151668 respectively) are in the Qwen tokenizer space. Pipeline B sliced trace and answer spans use these ids verbatim.

Model	Tokenizer	`</think>` id
R1-Distill-Qwen-7B	Qwen-2.5	151649
R1-Distill-Llama-8B	Llama-3 BPE	128014
QwQ-32B	Qwen-2.5	151668

Table 10: Tokenizer-specific `</think>` token ids used by Pipeline B (Section 3.3).