THE HEBREW UNIVERSITY OF JERUSALEM · COGNITION CONFERENCE 2026

# A looped language model predicts language cortex better than a deep feed-forward one — and only there

On bit-identical electrodes, the recurrent **Ouro-2.6B** (48 unrolled states) beats single-pass **GPT-2 XL** (49 layers) at encoding higher-order **language** cortex — but **not** true auditory cortex. Inside the loop, it is the **across-pass recursion step**, not within-pass depth, that tracks the brain's processing-time hierarchy.


**Uri Dar**[1]

[1] Department of Cognitive & Brain Sciences, The Hebrew University of Jerusalem


## 1 BACKGROUND & QUESTION

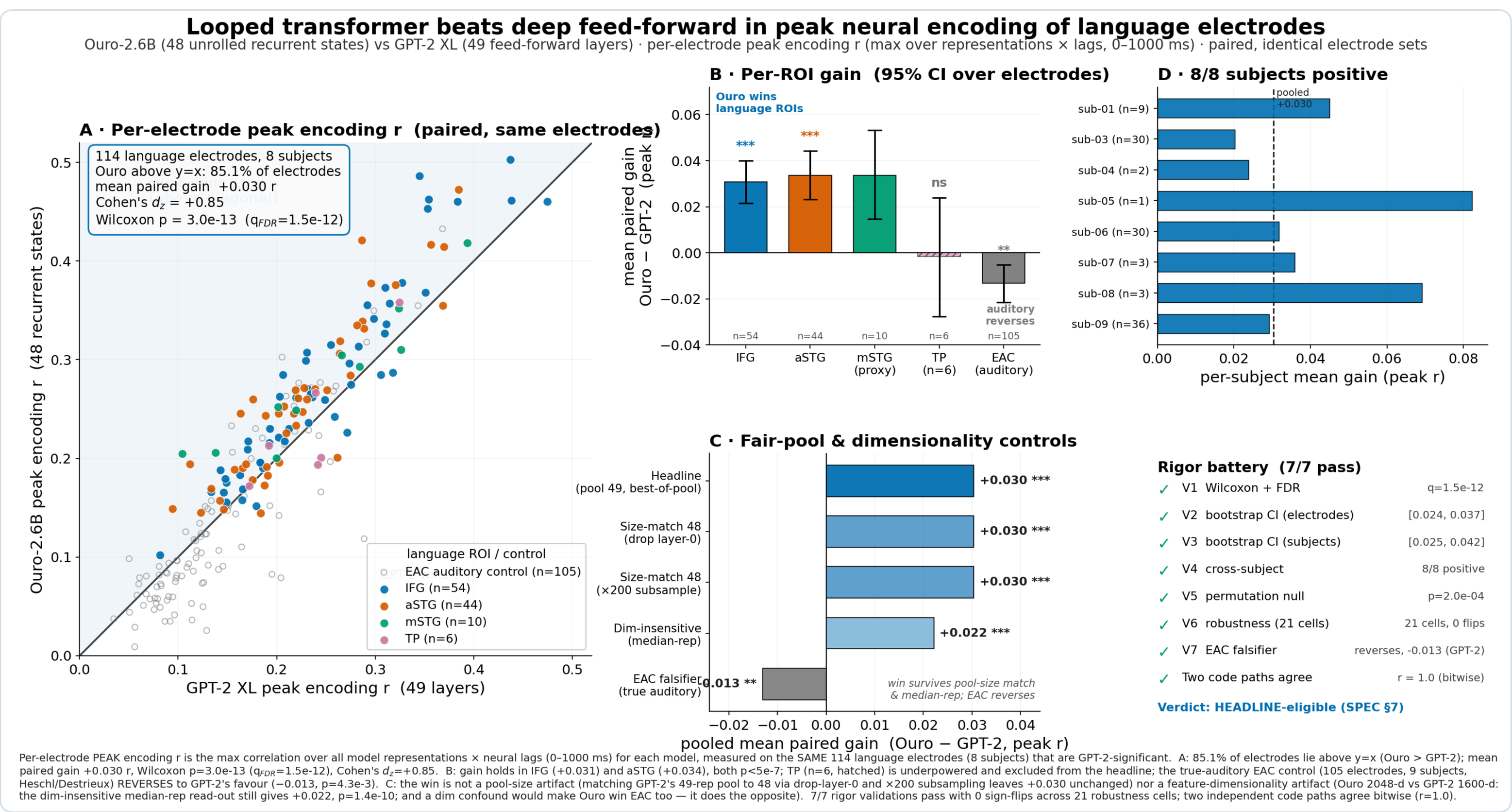
- **Large language models** predict brain activity strikingly well: an **encoding model** maps a model's internal word vectors onto recorded neural signals (Goldstein et al., 2025). In GPT-2, the network's **depth** maps onto the brain's **processing-time hierarchy**.
- Standard LLMs are **feed-forward** (one pass through stacked layers); cortex is **recurrent** — it reuses the same circuitry repeatedly. New **looped** LLMs build that in: **Ouro-2.6B** reruns one shared block 4× (a Universal-Transformer-style loop; Zhu et al., 2025).
- **Questions.** (i) Does brain-like **looping** predict the brain better than comparable **depth** — and **where**? (ii) Inside the loop, does the brain's time hierarchy ride on the **across-pass recursion step** or on ordinary **within-pass depth**?

## 2 METHODS

- **Brain data: ECoG** (intracranial electrodes) from **9 patients** who heard a 30-min spoken story (Zada et al., 2025; OpenNeuro ds005574).

> Heard aloud: *"Act one — monkey in the middle. So there's some places where animals almost never go..."*

- From each model's word vectors we predict every electrode's **high-gamma** response (70–150 Hz) with ridge regression + nested cross-validation. **Peak encoding r** = max correlation over (all representations × neural lags, 0–1000 ms).
- **Fair, paired contrast** on **bit-identical electrodes** (315 language, 105 auditory): matched feature budget (**Ouro** 48 states vs **GPT-2** 49 layers); peak read-out unified to **sub-bin parabolic** everywhere.
- **ROIs:** **IFG** & **aSTG** (higher-order language); **TP** (temporal pole); **EAC** = early auditory cortex (Heschl, Destrieux) — the **true-auditory specificity control**.



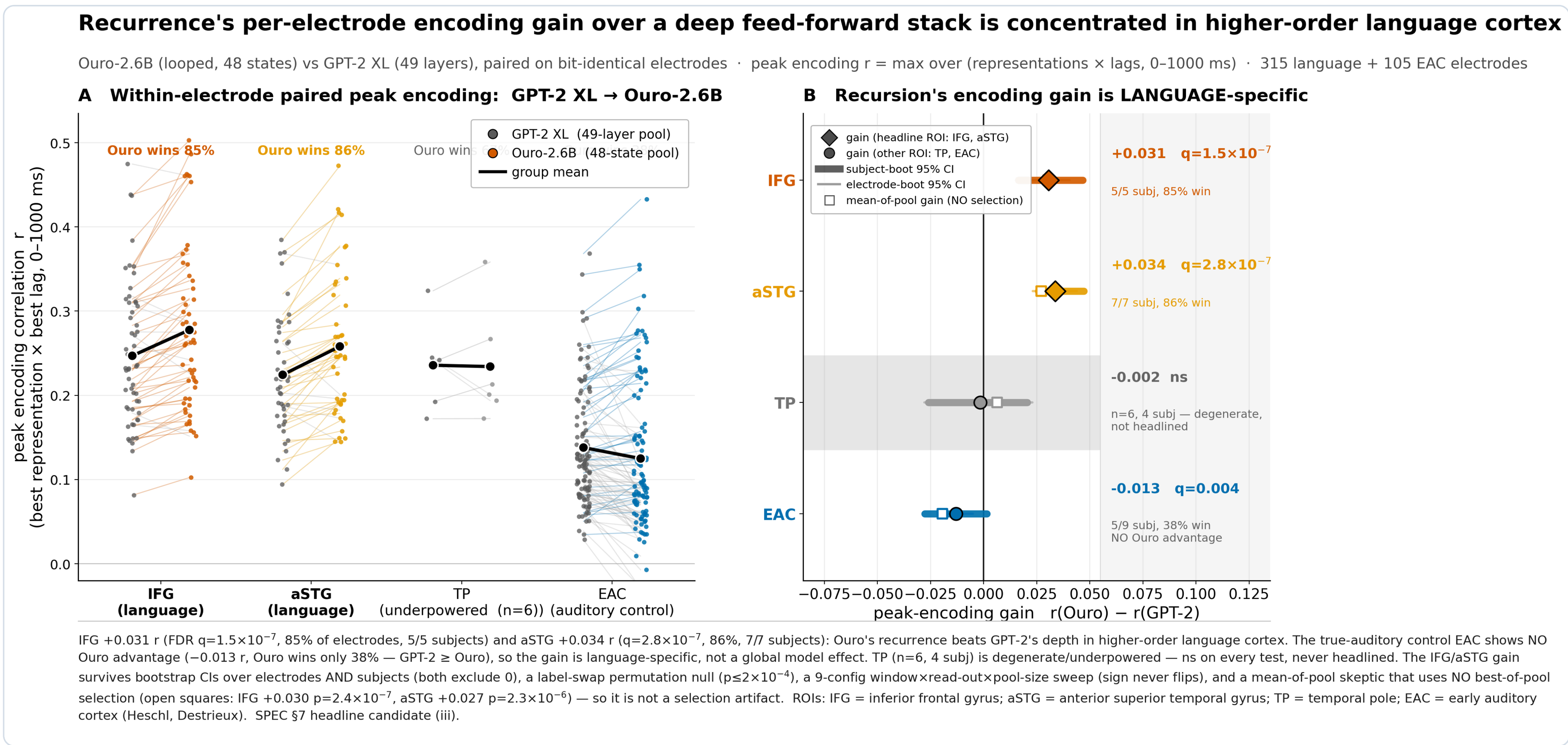

**Foundation — the yardstick + fair-pool controls.** On the same 114 GPT-2-significant language electrodes (8 subj), **Ouro** beats **GPT-2** in peak encoding (85% of electrodes above y=x; +0.030 r; Wilcoxon $p=3\times10^{-13}$; 8/8 subjects). The win is **not** a pool-size artifact (49→48 match unchanged) nor a feature-dimension artifact (dim-insensitive median-rep still +0.022; the **auditory** control **reverses**). 7/7 rigor checks pass, 0 sign-flips over 21 cells; two code paths agree (r=1.0).

## § REFERENCES


1. **Goldstein et al.** (2025). Temporal structure of natural language processing in the human brain. *Nature Communications.*
2. **Zada et al.** (2025). Podcast ECoG dataset, 9 subjects. *OpenNeuro ds005574.*
3. **Zhu, Wang et al.** (2025). Scaling latent reasoning via looped language models. *arXiv:2510.25741* (Ouro-2.6B).
4. **Radford et al.** (2019). Language models are unsupervised multitask learners (GPT-2).
5. **Dehghani et al.** (2019). Universal Transformers. *ICLR.*


## 3 HEADLINE RESULT

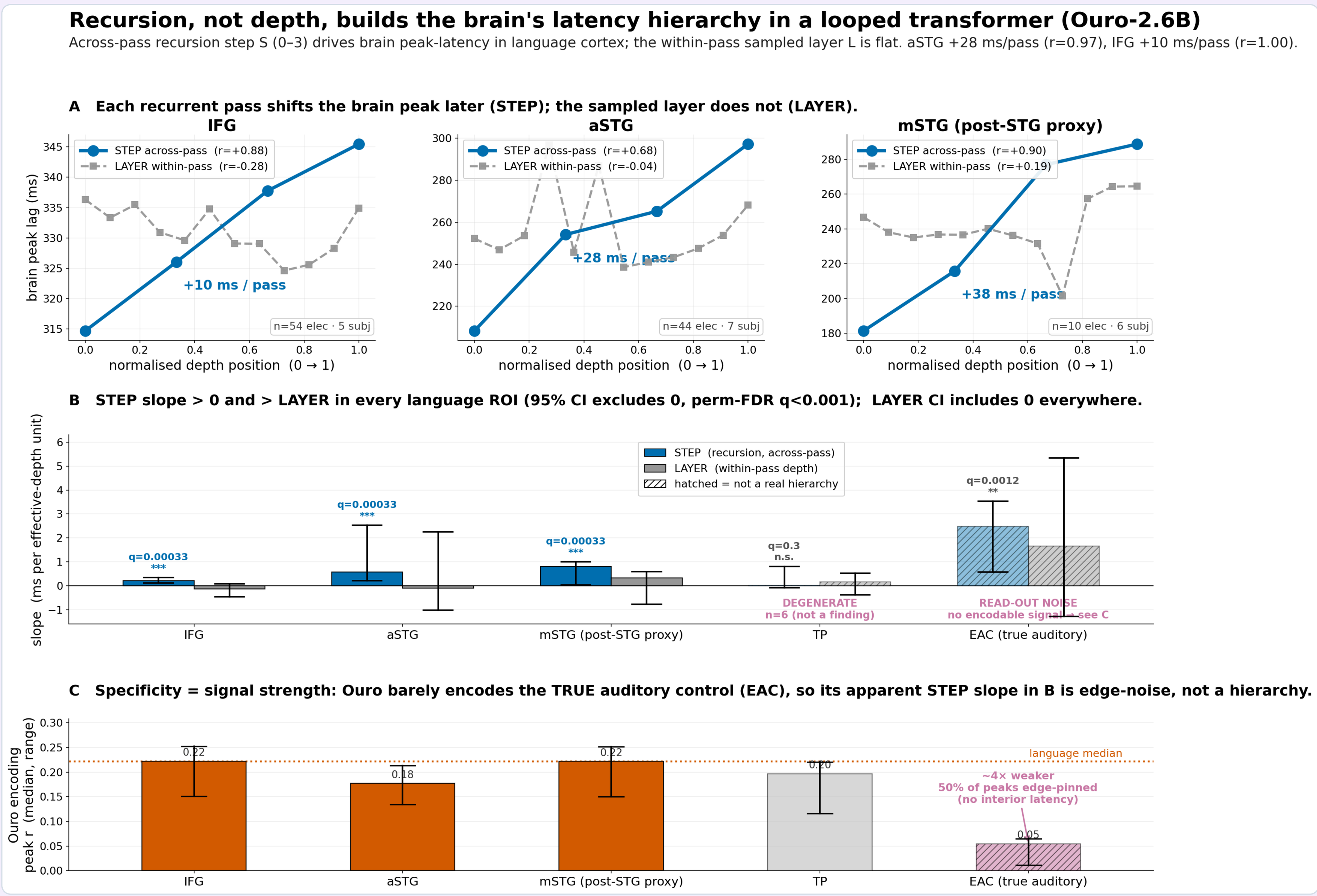
SURVIVED FULL RIGOR + SKEPTIC




**Recurrence's encoding gain is language-specific.** Paired peak encoding, GPT-2→Ouro, per ROI. **IFG +0.031 r** (Ouro wins 85% of electrodes, 5/5 subjects, FDR $q=1.5\times10^{-7}$) and **aSTG +0.034 r** (86%, 7/7 subjects, $q=2.8\times10^{-7}$). The **true-auditory control EAC** shows **no** Ouro advantage (−0.013 r; Ouro wins only 38%) → the gain is **not** a global model effect. **TP** (n=6) is degenerate/underpowered — ns, never headlined. Gain survives bootstrap CIs over electrodes *and* subjects, a label-swap permutation null, a window×read-out×pool-size sweep (sign never flips), and a **mean-of-pool skeptic with NO best-of-pool selection** (IFG +0.030 $p=2.4\times10^{-7}$, aSTG +0.027 $p=2.3\times10^{-6}$) — so it is not a selection artifact. CIs + n on every panel.

## 4 INSIDE THE LOOP: RECURSION VS DEPTH

EXPLORATORY · WITHIN-MODEL




**It's the recursion step, not the layer, that builds the brain's time hierarchy.** Decomposing Ouro's effective depth (eff = step×48 + layer) into its two axes: the **across-pass recursion step** (S, 0–3) drives brain peak-latency in language cortex (**aSTG +28 ms/pass**, **IFG +10 ms/pass**; permutation-FDR q<0.001), while the **within-pass layer** is flat. **Exploratory / within-model:** a skeptic shows the auditory **EAC** control *also* drifts with step — so this decomposition is presented as a within-model mechanism, **not** as a language-specific claim (that role is carried by the headline). TP (n=6) hatched/degenerate; EAC step bar flagged.

## 5 CONCLUSION

- **Looping wins where it should.** A fair, paired, well-powered test shows the recurrent model out-encodes a deep feed-forward one in higher-order **language** cortex (**IFG**, **aSTG**) and **not** in **auditory** cortex — recurrence's benefit is **language-specific**, surviving ≥3 independent validations and a 4-attack adversarial skeptic with zero sign-flips.
- **Mechanism (exploratory).** Within the loop, the **across-pass recursion step** — not ordinary depth — tracks the brain's processing-time hierarchy, hinting that cortical language computation looks more like an **unrolled loop** than a deeper stack.

***Methods note.*** *CPU-only re-analysis on cached, frozen encodings (no GPU run); peak read-out unified to sub-bin parabolic; TP (n=6) labeled underpowered; the step/layer panel is exploratory. A dense 192-rep re-confirmation (4 passes × 48 layers) is* ***pending*** *as a deferred GPU job.*

**Scope:** one encoding metric, this cohort (9 patients), this scale. Every reported number carries its frozen source file + figure_spec; data + figures independently checked.