ONE PASS, MANY VERDICTS: PARALLEL NEURAL DECISIONS ACROSS FOUR SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain activity is recorded through physically different windows: single-unit spikes, local field potentials, regional firing proportions and calcium fluorescence. It is interrogated with equally heterogeneous questions: which way will the animal choose, how fast will it react, was it rewarded, is it engaged. Today every (signal, question, session) triple receives its own decoder. We introduce **brainllax**, a *parallel decision model* for neural data that rebuilds the underlying method of the recently released Jev decision model and replaces this patchwork with a single interface. It takes neural activity from *any* subset of the four signals together with a batch of N typed judgement queries (categorical choices, ordinal scores and Booleans), and encodes a global neural state with a Perceiver-style encoder. Query-specific cross-attention and type-matched softmax, rank-consistent ordinal and sigmoid heads then return all N probability distributions in *one non-autoregressive pass*. We evaluate it on 39 Neuropixels sessions (10 mice, 73 brain regions) and brain-wide two-photon imaging, spanning 14 behavioural queries. One 12.7M-parameter model reaches mean Cohen's $\kappa = 0.660$, within 0.020 of 379 separately tuned per-session decoders, and a three-seed ensemble surpasses them (0.683) with the lowest calibration error of all methods. brainllax outperforms single-task transformers and pooled multi-task heads, answers from any of the 15 signal combinations with one set of weights, decides 64 queries $29.2\times$ faster than a comparably accurate but worse-calibrated autoregressive decoder, and adapts to unseen animals by re-learning only channel identities. The entire study ran in under four hours of wall-clock time on one two-GPU workstation.

1 INTRODUCTION

Systems neuroscience now records brain-wide activity through several physically distinct windows at once. High-density probes return single-unit *spikes* from hundreds of neurons across dozens of regions (Jun et al., 2017; Steinmetz et al., 2021); the same electrodes carry the *local field potential* (LFP), a mesoscale signal dominated by synaptic currents (Buzsáki et al., 2012; Einevoll et al., 2013); population summaries such as the *fraction of a region's neurons that fire* in a time bin approximate what mesoscale methods read out (Ma et al., 2016); and two-photon microscopy reports *calcium* fluorescence from tens of thousands of cells at a slower, nonlinear, indicator-limited timescale (Chen et al., 2013; Stringer et al., 2019b). The questions asked of these recordings are just as heterogeneous: *which way* will the mouse turn the wheel, *how fast* will it react, *was* the previous trial rewarded, *how dilated* is the pupil, *is* the animal running (Steinmetz et al., 2019; Musall et al., 2019; Urai et al., 2022).

The dominant practice answers each (signal, question, session) triple with its own decoder (Glaser et al., 2020). In the Neuropixels dataset studied here that amounts to hundreds of separately tuned models, none of which can exploit a second signal, a related question, or another animal. Recent neural foundation models push toward sharing across sessions (Azabou et al., 2023; Ye et al., 2023; Zhang et al., 2024; Azabou et al., 2025), but they are built around a single modality and typically read out one behavioural target per head. Generic multimodal transformers offer a different template: tokenise everything and *generate* the answer autoregressively (Reed et al., 2022; Lu et al., 2023). For a scientist interrogating a brain state this is the wrong interface: the answers are not text, they are

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.

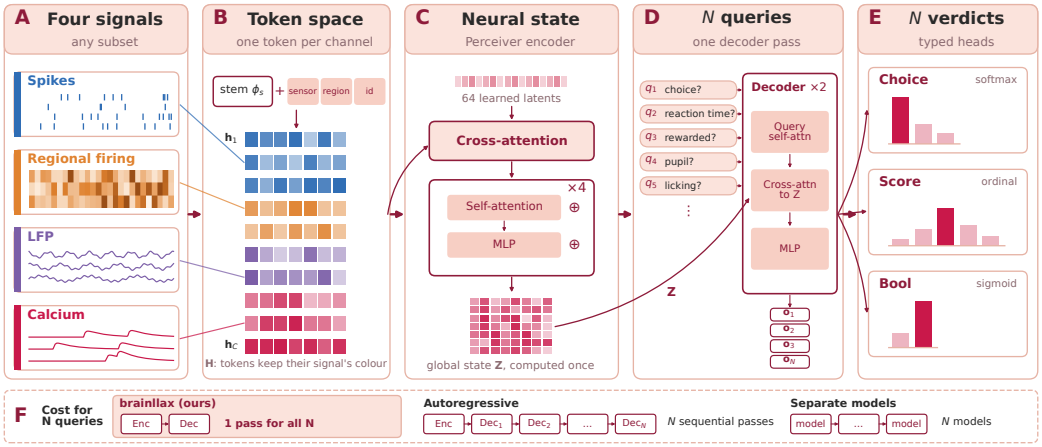

Figure 1: **brainllax at a glance.** (A) Any subset of four neural signals forms the context. (B) Every channel becomes one token through a sensor-specific stem plus shared sensor, region and channel-identity embeddings; tokens keep the colour of their signal. (C) A Perceiver-style Transformer (cross-attention from 64 learned latents, then four pre-norm self-attention blocks) distils all tokens into a global state $\mathbf{Z}$, computed once. (D) N typed queries are submitted together and decoded in one pass by query self-attention (no causal mask) and cross-attention to $\mathbf{Z}$. (E) Type-matched heads return a softmax (choice), a rank-consistent ordinal distribution (score) or a Bernoulli (bool) for every query. (F) Cost of answering N queries compared with autoregressive decoding and separate models.

distributions over a small set of outcomes, and there is no reason for the verdict on reaction time to wait for the verdict on choice.

We take the opposite route and build a *parallel decision model*. Our starting point is Jev, a recently released "System One" decision model (TypeSafe, 2026; Yu & Yao, 2026). Jev reads a state once and returns typed decisions (choices, scores and Boolean "Noul" judgements) with calibrated probabilities in a single non-autoregressive pass. brainllax rebuilds this underlying method for neural data. Its input is a *context state*, namely neural activity from any subset of the four signals, together with a *batch of N typed judgement queries*: categorical *choices*, ordinal *scores* and *Boolean* tests. A Perceiver-style encoder (Jaegle et al., 2021) compresses every channel of every signal into a fixed global state; a query decoder lets the N queries attend to that state and to each other; independent type-specific heads return all N probability distributions *in one forward pass* (Figure 1). No token is generated; no query waits on another.

Contributions.

- **A decision-native interface for neural data.** Following Jev's typed-decision paradigm, we formalise neural decoding as answering a *set* of typed queries against a shared context and give an architecture that answers any subset of queries in $O(1)$ decoder passes, with typed heads that respect the structure of each answer (Section 2).
- **One model for four signals.** To our knowledge this is the first model that consumes spikes, LFP, regional firing proportions and calcium through a single set of weights and remains usable when *any* of the 15 non-empty signal combinations is present at test time.
- **One network in place of 379 specialists.** On 34 Neuropixels sessions, 73 brain regions and a brain-wide calcium recording, one 12.7M-parameter brainllax reaches mean $\kappa = 0.660$ across 14 queries, within 0.020 of 379 separately tuned per-session linear decoders (0.680); an ensemble of three seeds surpasses them (0.683) and is the best-calibrated method in the study (ECE 0.038 vs. 0.069). brainllax outperforms 14 single-task transformers (0.627) and pooled multi-task heads (0.653, on every seed).
- **Speed, transfer and readout for free.** Parallel decoding answers 64 queries 29.2× faster than autoregressive decoding at near-equal accuracy and better calibration, and is immune to a label-

leakage failure that we identify in autoregressive decision decoders; on five unseen animals, re-learning only channel identities from 16 trials per session beats training from scratch by a wide margin; region occlusion yields a brain-wide map of which areas carry which verdict (Section 4).

The whole study, including preprocessing, ≈30 training runs and all analyses, fits in under four hours on two GPUs.

2 BRAINLLAX: PARALLEL DECISIONS OVER A UNIFIED NEURAL STATE

2.1 PROBLEM: A CONTEXT AND A SET OF TYPED QUERIES

A sample is a pair $(\mathcal{X}, \mathcal{Q})$. The *context* $\mathcal{X} = \{\mathcal{X}^m\}_{m\in\mathcal{M}}$ holds whatever signals were recorded, $\mathcal{M} \subseteq \{\text{SPK}, \text{RGN}, \text{LFP}, \text{CA}\}$, each a set of channels $\mathcal{X}^m = \{(\mathbf{x}_c, u_c, r_c)\}_c$ with a time series $\mathbf{x}_c \in \mathbb{R}^{L_m}$, a channel identity u_c (a unit, an electrode, a region summary) and an atlas region r_c. The *query set* $\mathcal{Q} = \{q_1, \dots, q_N\}$ lists the judgements requested for this context. Each query q has a type $\tau(q) \in \{\text{choice}, \text{score}, \text{bool}\}$ and an outcome space of size K_q. The model must return N distributions $p_\theta(y_q \mid \mathcal{X}, \mathcal{Q})$, one per query, *jointly and in one pass*. A query may be inapplicable to a context (reaction time on a no-go trial, reward on a passive trial); such *Null* labels are masked from the loss rather than forced into a class, so that the same query vocabulary serves every dataset.

2.2 ONE TOKENISER FOR FOUR SIGNALS

Every channel of every signal becomes one token,

$$\mathbf{h}_c = \phi_{s(c)}(\tilde{\mathbf{x}}_c) + \mathbf{e}^{\text{sensor}}_{s(c)} + \mathbf{e}^{\text{region}}_{r_c} + \mathbf{e}^{\text{id}}_{u_c}, \tag{1}$$

where ϕ_s is a two-layer MLP *stem* specific to the sensor s (Neuropixels spikes, regional proportions, LFP, simulated or two-photon calcium), $\tilde{\mathbf{x}}_c$ a light normalisation (square-root counts for spikes; per-session standardisation for LFP and calcium; LFP stems also receive six log band-powers), and three embeddings shared across all sensors: the sensor, one of 76 atlas regions, and a learnable channel identity in the spirit of POYO's unit embeddings (Azabou et al., 2023). Only the stems see sensor-specific sampling rates (50 Hz spike bins, 100 Hz LFP, 30 Hz simulated and ≈ 3 Hz two-photon calcium); from the embedding onwards all signals live in the same d-dimensional space. A learned session token is appended to the context.

2.3 GLOBAL STATE ENCODER

Channel counts vary from a handful (LFP) to thousands (calcium), so we use a latent bottleneck (Jaegle et al., 2021). 64 learned latents cross-attend to all channel tokens (padding masked), followed by 4 pre-norm self-attention blocks, yielding the global state $\mathbf{Z} \in \mathbb{R}^{64\times d}$. Cost is linear in the number of channels and independent of the number of queries.

2.4 PARALLEL DECISION DECODER AND TYPED HEADS

Each query is embedded as $\mathbf{q}_i = \mathbf{e}^{\text{task}}_{q_i} + \mathbf{e}^{\text{type}}_{\tau(q_i)}$. Two decoder blocks apply *query self-attention* (queries may exchange evidence), *query-specific cross-attention* into $\mathbf{Z}$ and an MLP; no causal mask is used and padding masks allow a different query set per sample. Output $\mathbf{o}_i$ is read by an independent linear head W_{q_i} whose parametrisation follows the answer type:

$$\text{choice: } p(y{=}k) = \text{softmax}_k(W_q\mathbf{o})_k, \qquad \text{bool: } p(y{=}1) = \sigma(\mathbf{w}_q^\top \mathbf{o}), \tag{2}$$

$$\text{score: } p(y > k) = \sigma\big(\mathbf{w}_q^\top \mathbf{o} - \textstyle\sum_{j\le k} \text{softplus}(b_{q,j})\big), \quad k = 0, \dots, K_q - 2, \tag{3}$$

the last being a rank-consistent ordinal (CORAL-style, Cao et al., 2020) head whose cumulative thresholds are monotone by construction, so that class probabilities $p(y{=}k) = p(y{>}k{-}1) - p(y{>}k)$ are always valid.

Proposition 1 (Decision cost)**.** *Let C_{enc} be the encoder cost and $C_{dec}(n)$ the cost of one decoder pass over n queries. brainllax answers N queries in $C_{enc} + C_{dec}(N)$ with one sequential step. An autoregressive decoder that conditions query i on answers $1{:}i{-}1$ requires N sequential steps and, without caching, $C_{enc} + \sum_{i=1}^{N} C_{dec}(i) = C_{enc} + \Theta(N^2)\, C_{dec}(1)$; answering with N separate models costs $N(C_{enc} + C_{dec}(1))$.*

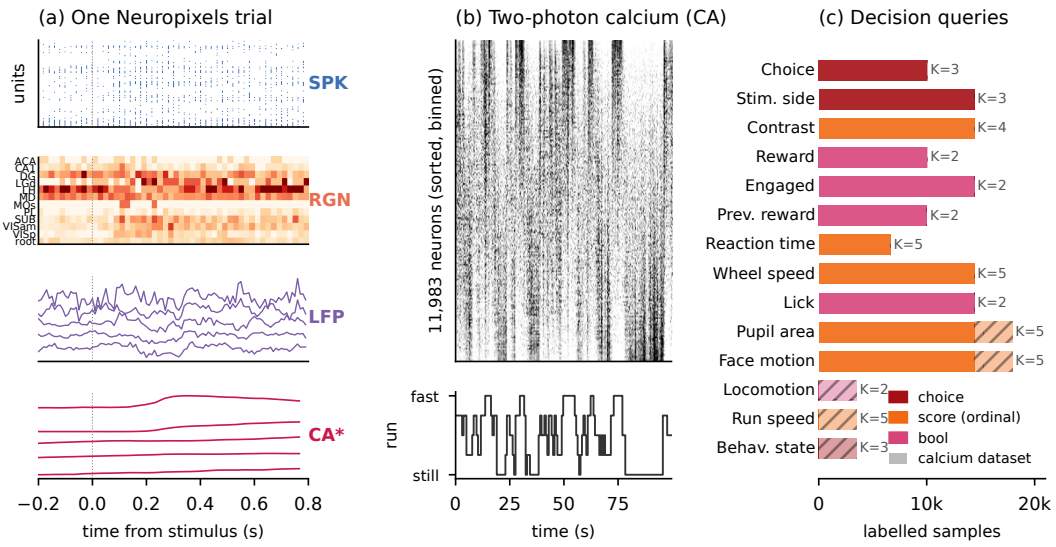

Figure 2: **A unified four-signal decision corpus.** (a) One trial of the visual-discrimination task: spike raster of the units observed as spikes (SPK), regional firing proportion for each recorded area (RGN), LFP (LFP) and simulated GCaMP6s fluorescence of the complementary units (CA*); dotted line = stimulus onset. (b) Brain-wide two-photon calcium from 11,983 neurons (sorted by first principal component, binned) with the running-speed score that the model must judge. (c) The 14 decision queries, coloured by type; hatched = labels from the calcium dataset; K = number of outcomes.

Because $C_{\text{dec}}(N)$ is a small attention over $N{+}64$ tokens, on a GPU the parallel pass is effectively constant in N (Section 4.3). The proof is immediate and given in Section E.

2.5 Training

The loss is the mean over *answerable* queries of the type-appropriate negative log-likelihood, each divided by $\log K_q$ so that binary, ternary and five-level queries contribute on the same scale. Two forms of dropout make the interface robust. *Modality dropout* removes each available signal with probability 0.3 (keeping at least one), so a single network learns to answer from any signal subset. *Query-set dropout* withholds each query with probability 0.2, so the decoder never relies on a fixed query set. Training additionally sub-samples up to 1,024 channels per signal per sample. We use AdamW (Loshchilov & Hutter, 2019), learning rate 5×10^{-4} with warm-up and cosine decay, batch 64, 50 epochs, bf16; the model selected on validation κ is evaluated once on the test split. Full hyper-parameters are in Section D.

3 Experimental Setup

Data. We build a corpus from two public mouse datasets (Figure 2; details in Section B). *Neuropixels*: 39 sessions from 10 mice performing a two-alternative contrast discrimination task with a no-go option, plus passive stimulus blocks (Steinmetz et al., 2019); 14,420 trials, 33,958 units in 73 atlas regions and 4–18 LFP channels per session. Per trial we use a 1 s window (-0.2 to $+0.8$ s around stimulus onset). The units of each session are split at random into two disjoint halves. One half is observed as spike counts (SPK) and the other through a GCaMP6s forward model (CA*; rise 50 ms, decay 0.7 s, log-normal gain, Gaussian noise, 30 Hz; Section C), mimicking a preparation where electrophysiology and imaging report different cells. RGN tokens give, per region and 20 ms bin, the fraction of that region's units that fired. *Two-photon*: the brain-wide spontaneous-behaviour recording of Stringer et al. (2019b) (11,983 neurons, deconvolved, ≈ 3 Hz), windowed into 3,504 overlapping 10-frame contexts. Five Neuropixels sessions (one each from five mice) are held out entirely for transfer experiments; the remaining 34 sessions are split 70/10/20 by trial and the calcium recording into contiguous temporal blocks with guard gaps.

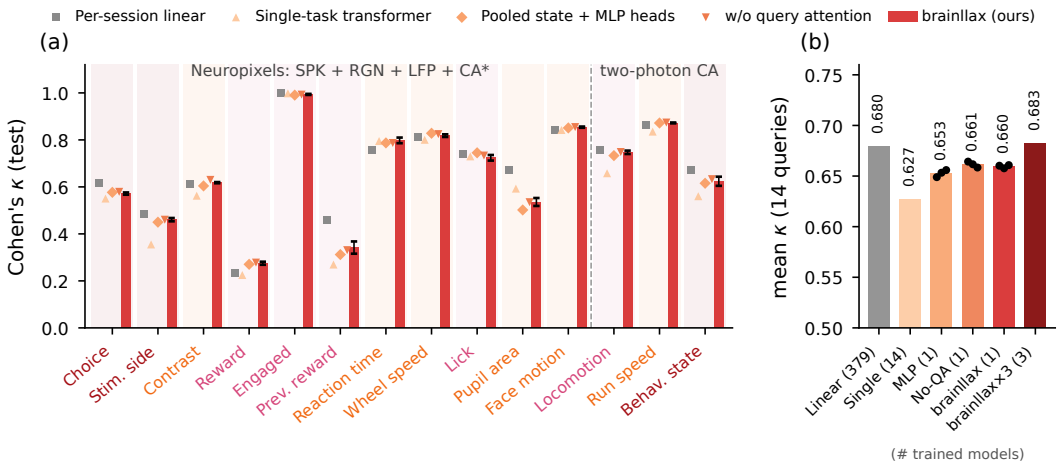

Figure 3: **Main results.** (a) Test κ for each of the 14 queries (background colour = query type). Bars: brainllax (mean $\pm$ s.d. over three seeds); markers: baselines. Left of the dashed line: Neuropixels queries answered from four signals; right: two-photon queries. (b) Mean κ over all queries; dots are seeds; numbers in brackets give how many models each approach trains. brainllax$\times$3 averages the probabilities of the three seeds.

Queries. Fourteen queries span the three types (Figure 2c): *choice* (turn left / no-go / right), *stimulus side*, *behavioural state* (quiescent / whisking / running); *scores* for stimulus contrast, reaction time, wheel speed, pupil area, face motion and running speed (quantile bins per session); and *bools* for reward, engagement (task vs. passive block), previous-trial reward, licking and locomotion. Pupil area and face motion are *shared* queries answered from both datasets with the same query embedding.

Metric. Our headline metric is Cohen's κ on the test split (Cohen, 1968), unweighted for choice and bool and quadratic-weighted for ordinal scores. It is 0 at chance and 1 at perfect agreement for every query regardless of class balance, and so can be averaged across queries. We additionally report balanced accuracy, AUROC, Spearman ρ, negative log-likelihood and expected calibration error (ECE, 10 bins; Naeini et al., 2015; Guo et al., 2017).

Baselines. (i) *Per-session linear*: the standard systems-neuroscience decoder (Steinmetz et al., 2019; Glaser et al., 2020), i.e. one L2-regularised logistic regression per session and query, on time-binned rates of all four signals (PCA to 256 dimensions), with the regularisation strength selected on validation: 379 models in total. (ii) *Single-task*: the brainllax architecture trained separately for each query (14 models). (iii) *MLP heads*: the same encoder with the global state mean-pooled and read by per-query heads, i.e. hard parameter sharing (Caruana, 1997; Ruder, 2017). (iv) *w/o query attention*: brainllax with query self-attention masked to the identity. (v) *Autoregressive*: brainllax with a causal decoder in which query i additionally receives an embedding of the answer to query $i-1$ (teacher-forced in training, own predictions at test), the decision-model analogue of token generation. (vi) *Specialists*: brainllax trained on a single signal. All transformer variants share the encoder, optimiser and budget; brainllax, (iii) and (iv) use three seeds.

4 Results

4.1 One network in place of hundreds of specialist decoders

Figure 3 and Table 1 summarise the central result. A single brainllax network reaches mean $\kappa = 0.660 \pm 0.001$ over the 14 queries. The standard systems-neuroscience approach trains a separately regularised linear decoder for every session and query. These 379 models, each of which sees only its own session, reach 0.680. brainllax is ahead on 6 of the 14 queries, most clearly on reaction time, reward, face motion and running speed. It trails on queries dominated by slow within-session drift

Table 1: **Test κ per query** (C = choice, S = score, B = bool). brainllax: mean $\pm$ s.d. over three seeds; MLP heads and w/o q-attn: mean of three seeds; others: one run. Bold: best per row.

Query	Type	Linear (379 models)	Single-task (14 models)	MLP heads (1 model)	w/o q-attn (1 model)	AR (1 model)	**brainllax** (1 model)
Choice	C	**0.615**	0.550	0.577	0.579	0.608	0.572 $\pm$0.005
Stim. side	C	0.486	0.355	0.450	0.460	**0.488**	0.461 $\pm$0.007
Contrast	S	0.611	0.562	0.604	0.630	**0.633**	0.618 $\pm$0.003
Reward	B	0.232	0.225	0.271	0.278	**0.289**	0.275 $\pm$0.006
Engaged	B	**1.000**	0.998	0.991	0.993	0.991	0.994 $\pm$0.002
Prev. reward	B	**0.460**	0.269	0.312	0.331	0.427	0.342 $\pm$0.026
Reaction time	S	0.756	0.795	0.787	0.787	**0.804**	0.798 $\pm$0.012
Wheel speed	S	0.811	0.800	**0.828**	0.825	0.823	0.818 $\pm$0.006
Lick	B	0.740	0.730	**0.745**	0.733	0.732	0.724 $\pm$0.012
Pupil area	S	**0.672**	0.593	0.502	0.535	0.585	0.536 $\pm$0.017
Face motion	S	0.842	0.842	0.851	0.855	**0.858**	0.854 $\pm$0.002
Locomotion	B	**0.755**	0.658	0.733	0.748	0.738	0.747 $\pm$0.007
Run speed	S	0.862	0.835	0.872	**0.874**	0.871	0.872 $\pm$0.002
Behav. state	C	**0.671**	0.560	0.615	0.633	0.608	0.624 $\pm$0.019
Mean (14)		**0.680**	0.627	0.653	0.661	0.675	0.660 $\pm$0.001

(pupil area, previous reward), which a per-session model can absorb into its intercept. Averaging three seeds closes the gap entirely: the ensemble reaches 0.683, above the 379 specialists. Training the same transformer once per query (14 single-task models) reaches 0.627; brainllax is ahead on 11 of 14 queries, so answering queries jointly is a gain rather than a compromise. Conventional hard parameter sharing uses the same encoder but reads a mean-pooled state with per-query heads. brainllax improves on it for every seed (0.660 vs. 0.653): letting each query *attend* to the global state is better than giving all queries the same pooled summary. Masking query-to-query attention leaves accuracy unchanged within seed noise (0.661); the benefit of the decision decoder comes from query-specific cross-attention, and query interaction is a free, interpretable by-product (Section 4.3). Calibration follows the same pattern: a single brainllax has lower NLL than the per-session decoders (0.720 vs. 0.731) and the three-seed ensemble has the lowest ECE of any method (0.038; Table 4).

4.2 Any signal, any subset, one set of weights

Because every signal lives in the same token space, the trained model can be queried with any of the 15 non-empty signal combinations (Figure 4a, Table 5). Performance degrades gracefully, from 0.631 with all four signals to 0.483 on average with a single one, and every combination remains well above chance. Fusing all four signals beats the best single signal on 10 of 11 verdicts (Figure 4c). The signals are complementary: SPK and CA* report disjoint neurons and their combination is the strongest pair; RGN summarises whole areas in a handful of tokens yet alone supports most queries; LFP alone carries movement information (wheel speed, face motion, reaction time). Restricted to the information-rich signals (Figure 4b), the unified model is as good as a specialist trained on that signal alone (SPK: 0.604 vs. 0.600; CA: 0.603 vs. 0.603), i.e. co-training on four signals does not dilute any of them. For the low-bandwidth signals seen alone only rarely during training, specialists remain better (RGN: 0.497 vs. 0.574; LFP: 0.226 vs. 0.398), pointing to modality-dropout schedules as a lever for future work. Modality dropout is what makes any-subset use possible at all: without it the model falls to 0.153 on average from one signal (Figure 4b). It is also a strong regulariser when all signals are present: over all 14 queries the model trained without it reaches only 0.625, against 0.660 with it.

4.3 Parallel decisions: faster, better calibrated, and leak-free

Latency is where the decision-model view pays off most directly (Figure 5a,b). Because the global state is computed once and all queries are decoded jointly, brainllax's latency is almost flat in N: 5.44 ms for one query and 5.43 ms for 64 queries on a single context. An autoregressive decoder needs 158 ms for the same 64 queries (29.2$\times$ slower) and 64 separate models 341 ms (62.8$\times$). At batch 32 and 128 queries the gaps are 58.3$\times$ and 125$\times$, and brainllax delivers about 7×10^5

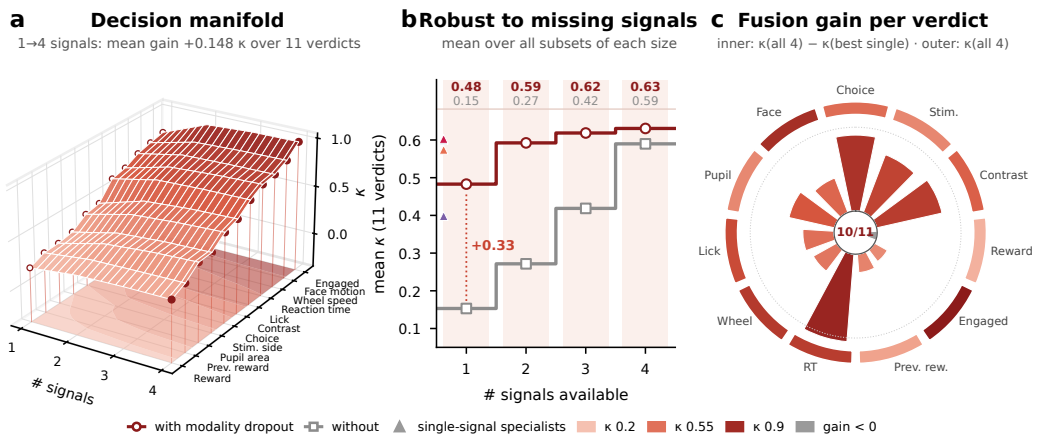

Figure 4: **Signal-agnostic decisions.** (a) Decision manifold: test κ of one trained brainllax for each Neuropixels verdict (sorted by full-context κ) as a function of the number of available signals, averaged over all subsets of that size and three seeds; open and filled dots mark one and four signals, and the floor shows the contour projection. (b) Mean κ over the 11 verdicts per subset size (steps and markers; values listed along the top, red with and grey without modality dropout); triangles mark specialists trained on a single signal. (c) Fusion gain per verdict: inner bars show κ with all four signals minus κ with the best single signal, (dotted ring: $+0.04$), and the outer ring encodes κ with all four signals. Full values are in Table 5.

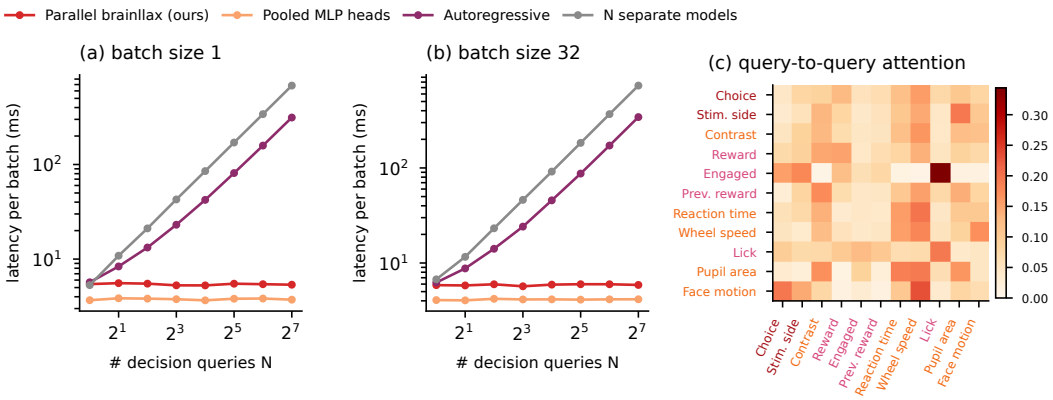

Figure 5: **Parallel decision decoding.** (a, b) GPU latency to answer N queries for one context and for a batch of 32 contexts (log–log). brainllax answers all queries in one decoder pass; the autoregressive variant needs N sequential passes; separate models need N encoder passes. (c) First-layer query self-attention of brainllax, averaged over Neuropixels test trials (rows attend to columns).

decisions per second on one GPU. What does sequential conditioning buy in return? A correctly trained autoregressive decoder, which sees its own previous verdicts, reaches $\kappa = 0.675$ (one seed) versus 0.660. The small gain is concentrated on logically coupled queries (choice after stimulus side, previous reward), and it is paid for twice. First, the probabilities are markedly worse (NLL 0.938 vs. 0.720, ECE 0.113 vs. 0.080; Table 4), because committing to hard answers turns every early mistake into confidently wrong context for later queries. Second, latency grows linearly with N. Autoregression also carries a subtle risk that parallel decoding avoids by construction: when inapplicable (Null) answers are fed back during teacher forcing they leak information about other queries, and a naively trained autoregressive decoder collapses on engagement ($\kappa = 0.000$; Section F.2). Parallel decoding thus sits on the favourable side of an accuracy–calibration–latency trade-off: near-identical accuracy, better-calibrated distributions, constant latency. The query self-

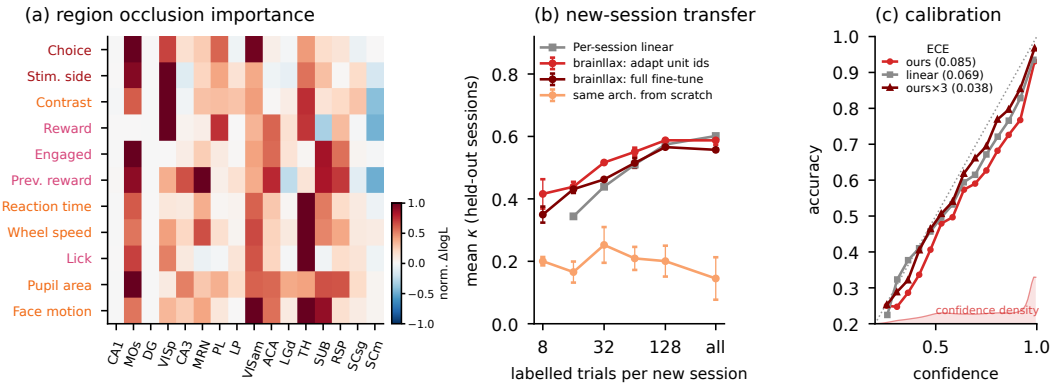

Figure 6: **Readout, transfer and calibration.** (a) Region-occlusion importance: drop in test log-likelihood when all tokens from one region (all signals) are removed, normalised per query; 16 most-sampled regions. (b) Five entirely held-out sessions (five mice): mean κ vs. number of labelled trials per session for adapting only the channel-identity and session embeddings of the frozen brainllax, full fine-tuning, training the same architecture from scratch, and per-session linear decoders (mean $\pm$ s.d. over two draws). (c) Reliability diagram pooled over all queries (18 confidence bins); the shaded band at the bottom is the density of the model's confidence on test predictions.

attention of brainllax is far from diagonal (Figure 5c): queries consult other queries, and the sharpest couplings are behaviourally meaningful. Engagement attends to licking, face motion to wheel speed and choice to reaction time, giving a learned, data-driven map of which verdicts travel together.

4.4 What the model reads, how it transfers, and how much to trust it

A brain-wide map of verdicts. Occluding one region at a time (Figure 6a, Figure 11) yields a query-by-region importance map from a single trained model. Evidence for most queries is spread over many areas, in line with the distributed coding of choice, action and engagement reported by Steinmetz et al. (2019). Among the 16 most-sampled regions, however, the peaks are anatomically sensible: stimulus side and contrast rely most on primary visual cortex (VISp); movement-linked verdicts (reaction time, wheel speed, licking, face motion) on the thalamus and secondary motor cortex (MOs); choice on MOs and anteromedial visual cortex; engagement and pupil on MOs and the subiculum. The map is a model-based hypothesis generator rather than a causal claim, but it is obtained for all queries simultaneously and at no extra training cost.

New animals from a handful of trials. For five sessions from five mice never seen in training, we freeze brainllax and learn only the new channel-identity and session embeddings (Figure 6b). With 16 labelled trials per session this reaches $\kappa = 0.439$, versus 0.166 for the same architecture trained from scratch and 0.344 for per-session linear decoders. The advantage over linear decoders persists up to 128 trials (0.588 vs. 0.574), and with all $\approx$210 training trials per session the two are on par (0.587 vs. 0.601). Adapting only identities is consistently better than full fine-tuning (0.431 at 16 trials), which overfits. The shared decision machinery transfers across animals: in the low-data regime that matters for new recordings, a session needs only to be *introduced* to the model, not re-learned.

Calibrated verdicts. Because every head is a proper likelihood matched to its answer type, brainllax's confidence tracks its accuracy across all three query types (Figure 6c, Figure 12). A single model is mildly over-confident (ECE 0.080), as is typical of trained transformers (Guo et al., 2017). Averaging three seeds removes this almost entirely (ECE 0.038, the best of all methods), making the decoded verdicts usable as probabilistic inputs to downstream analyses or closed-loop experiments.

5 Related work

Neural population decoders and foundation models. Latent-dynamics models (Pandarinath et al., 2018; Vyas et al., 2020) and contrastive embeddings (Schneider et al., 2023) relate population activity to behaviour, and transformers now model spiking data directly (Ye & Pandarinath, 2021; Le & Shlizerman, 2022; Pei et al., 2021). Multi-session pre-training scales these ideas: NDT2 learns context tokens across sessions and subjects (Ye et al., 2023), POYO tokenises individual spikes with learnable unit embeddings and a Perceiver backbone (Azabou et al., 2023), POYO+ adds calcium imaging from many cell types and multiple regression tasks (Azabou et al., 2025), multi-task masking learns a universal spike translator (Zhang et al., 2024), and Neuroformer generates spikes and behaviour autoregressively (Antoniades et al., 2024). EEG and intracranial models pursue the same goal for field signals (Wang et al., 2023; Yang et al., 2023; Zhang et al., 2023; Jiang et al., 2024). brainllax borrows the Perceiver bottleneck and unit identities from this line, but differs in two respects that we believe are essential: it places *electrophysiological, mesoscale and optical* signals in one token space, and it treats the *output side* as a first-class object, namely a variable set of typed judgement queries answered jointly.

Query-based and non-autoregressive prediction. Learned queries that attend into a latent state underlie DETR (Carion et al., 2020), Perceiver IO (Jaegle et al., 2022), the Q-Former (Li et al., 2023) and set transformers (Lee et al., 2019); non-autoregressive decoding trades sequential generation for parallel prediction (Gu et al., 2018). Generalist models instead serialise every output as tokens (Reed et al., 2022; Lu et al., 2023). Closest to our work, Jev evaluates many typed questions against one shared state in a single parallel pass, trained with reinforcement learning for calibrated decisions (TypeSafe, 2026), and Visual Jev extends the paradigm to shared visual context (Yu & Yao, 2026). brainllax rebuilds this parallel typed-decision method for neural recordings, with multi-signal tokenisation, Null-masking and a supervised likelihood objective in place of RL-based calibration. We show that for decision-style outputs the parallel route gives up little accuracy, improves calibration, and gains more than an order of magnitude in latency.

Multi-task learning. Hard parameter sharing (Caruana, 1997), task-weighting (Kendall et al., 2018; Chen et al., 2018), gradient surgery (Yu et al., 2020), task attention (Liu et al., 2019) and task grouping (Standley et al., 2020) address interference between heads. Our query decoder offers a complementary mechanism: each task reads the shared state through its own cross-attention, which in our experiments consistently beats a shared pooled representation, while query self-attention exposes task relations at inference time. An extended discussion, including cross-modal studies of calcium versus electrophysiology (Wei et al., 2020; Siegle et al., 2021b), is in Section A.

6 Discussion and limitations

brainllax shows that a single network, trained in minutes, can stand in for hundreds of per-session decoders across four kinds of neural signal, return calibrated distributions for a whole battery of questions in one pass, and remain useful when signals go missing or a new session arrives. The decision-model view also changes how neural data can be *queried*: adding a question means adding a query embedding, not training a new decoder, and each question reads the shared state through its own attention.

We are explicit about the boundaries of the present evidence. **(i) Simulated calcium.** On Neuropixels sessions the calcium signal is produced by a GCaMP6s forward model applied to real spikes of held-apart units; only the Stringer recording is genuine two-photon data. Our claims about calcium therefore rest on one real session plus a physiologically parameterised simulation; simultaneous multi-sensor datasets are the natural next test. **(ii) Decoding, not forecasting.** Queries such as wheel speed or face motion are judged from activity in the same window; they measure how much behaviourally relevant information the state carries, not predictive power ahead of movement. **(iii) Easy and hard queries.** Engagement is near-perfectly decodable because passive blocks follow task blocks, and some session-level drifts (pupil, previous reward) remain better captured by per-session linear models, which see each session's slow statistics directly. **(iv) Scale.** We deliberately worked at small scale (12.7M parameters, under four hours of wall-clock time on two GPUs). Whether the parity with per-session decoders becomes dominance with more sessions, as in POYO-style scaling

studies, is untested here. **(v) Single model vs. specialists.** A single brainllax is slightly below the per-session decoders on average (0.660 vs. 0.680), and only the seed ensemble overtakes them; query self-attention, although interpretable, does not by itself improve accuracy at this scale.

ETHICS STATEMENT

This work re-analyses two publicly released mouse datasets (Steinmetz et al., 2019; Stringer et al., 2019b); all animal procedures were approved and reported by the original studies, and no new animal experiments were performed. Decoding behaviour from neural activity has potential future relevance for brain–computer interfaces (Willett et al., 2021); our models are trained and evaluated on mice only and make no clinical claims.

REPRODUCIBILITY STATEMENT

Both datasets are publicly available. Sections B to D give every preprocessing step, the calcium forward model, all hyper-parameters, the splits (fixed random seed per session; contiguous blocks with guard gaps for calcium) and the evaluation protocol. The full pipeline, including preprocessing, ≈ 30 training runs, baselines and analyses, ran in under four hours on two NVIDIA RTX A6000 GPUs; per-run compute is listed in Table 6. Code, configuration files and the exact job list will be released upon publication.

STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

Large language models were used extensively in this work: to write and debug the experiment code, to draft and edit the manuscript text, and to produce figure scripts. All experiments were executed on real public data, every reported number is generated programmatically from the logged results, and the authors take full responsibility for the content, including the verification of all references.

REFERENCES

Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and multitask generative pretraining for brain data. In *International Conference on Learning Representations*, 2024.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael J. Mendelson, Blake Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mehdi Azabou, Krystal Xuejing Pan, Vinam Arora, Ian Jarratt Knight, Eva L. Dyer, and Blake Aaron Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *International Conference on Learning Representations*, 2025.

György Buzsáki, Costas A. Anastassiou, and Christof Koch. The origin of extracellular fields and currents: EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, 13(6):407–420, 2012.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229, 2020.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Tsai-Wen Chen, Trevor J. Wardill, Yi Sun, Stefan R. Pulver, Sabine L. Renninger, Amy Baohan, Eric R. Schreiter, Rex A. Kerr, Michael B. Orger, Vivek Jayaraman, Loren L. Looger, Karel Svoboda, and Douglas S. Kim. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300, 2013.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pp. 794–803, 2018.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

Saskia E. J. de Vries, Jerome A. Lecoq, Michael A. Buice, Peter A. Groblewski, Gabriel K. Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23(1):138–151, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Gaute T. Einevoll, Christoph Kayser, Nikos K. Logothetis, and Stefano Panzeri. Modelling and analysis of local field potentials for studying the function of cortical circuits. *Nature Reviews Neuroscience*, 14(11):770–785, 2013.

Johannes Friedrich, Pengcheng Zhou, and Liam Paninski. Fast online deconvolution of calcium imaging data. *PLOS Computational Biology*, 13(3):e1005423, 2017.

Juan A. Gallego, Matthew G. Perich, Raeed H. Chowdhury, Sara A. Solla, and Lee E. Miller. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, 23(2):260–270, 2020.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Joshua I. Glaser, Ari S. Benjamin, Raeed H. Chowdhury, Matthew G. Perich, Lee E. Miller, and Konrad P. Kording. Machine learning for neural decoding. *eNeuro*, 7(4), 2020.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664, 2021.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

James J. Jun, Nicholas A. Steinmetz, Joshua H. Siegle, Daniel J. Denman, Marius Bauza, Brian Barbarits, Albert K. Lee, Costas A. Anastassiou, Alexandru Andrei, Çağatay Aydın, et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Trung Le and Eli Shlizerman. STNDT: Modeling neural population activity with spatiotemporal transformers. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023.

Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1871–1880, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *International Conference on Learning Representations*, 2023.

Ying Ma, Mohammed A. Shaik, Mariel G. Kozberg, Sharon H. Kim, Jacob P. Portes, Dmitriy Timerman, and Elizabeth M. C. Hillman. Wide-field optical mapping of neural activity and brain haemodynamics: considerations and novel approaches. *Philosophical Transactions of the Royal Society B*, 371(1705):20150360, 2016.

Ulrich Mitzdorf. Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. *Physiological Reviews*, 65(1):37–100, 1985.

Simon Musall, Matthew T. Kaufman, Ashley L. Juavinett, Steven Gluf, and Anne K. Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, 22 (10):1677–1686, 2019.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Marius Pachitariu, Carsen Stringer, and Kenneth D. Harris. Robustness of spike deconvolution for neuronal calcium imaging. *Journal of Neuroscience*, 38(37):7976–7985, 2018.

Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815, 2018.

Felix Pei, Joel Ye, David Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. In *arXiv preprint arXiv:1706.05098*, 2017.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.

Joshua H. Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K. Ramirez, Hannah Choi, Jennifer A. Luviano, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021a.

Joshua H. Siegle, Peter Ledochowitsch, Xiaoxuan Jia, Daniel J. Millman, Gabriel K. Ocker, Shiella Caldejon, Linzy Casal, Andy Cho, Daniel J. Denman, Séverine Durand, et al. Reconciling functional differences in populations of neurons recorded with two-photon imaging and electrophysiology. *eLife*, 10:e69068, 2021b.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pp. 9120–9132, 2020.

Nicholas A. Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D. Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019.

Nicholas A. Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539): eabf4588, 2021.

Ian H. Stevenson and Konrad P. Kording. How advances in neural recording affect data analysis. *Nature Neuroscience*, 14(2):139–142, 2011.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019a.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019b.

The International Brain Laboratory. Standardized and reproducible measurement of decision-making in mice. *eLife*, 10:e63711, 2021.

Lucas Theis, Philipp Berens, Emmanouil Froudarakis, Jacob Reimer, Miroslav Román Rosón, Tom Baden, Thomas Euler, Andreas S. Tolias, and Matthias Bethge. Benchmarking spike rate inference in population calcium imaging. *Neuron*, 90(3):471–482, 2016.

TypeSafe. Introducing system one models and Jev. https://typesafe.ai/blog/introducing-system-one-models-and-jev, 2026. Model announcement, September 15, 2026.

Anne E. Urai, Brent Doiron, Andrew M. Leifer, and Anne K. Churchland. Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, 25(1):11–19, 2022.

Saurabh Vyas, Matthew D. Golub, David Sussillo, and Krishna V. Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations*, 2023.

Ziqiang Wei, Bei-Jung Lin, Tsai-Wen Chen, Kayvon Daie, Karel Svoboda, and Shaul Druckmann. A comparison of neuronal population dynamics measured with calcium imaging and electrophysiology. *PLOS Computational Biology*, 16(9):e1008198, 2020.

Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858): 249–254, 2021.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with Neural Data Transformers. *Neurons, Behavior, Data analysis, and Theory*, 2021.

Joel Ye, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: Multi-context pretraining for neural spiking activity. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Guanxu Yu and Yuhang Yao. Visual Jev: Accurate and efficient decisions from shared visual context. *arXiv preprint arXiv:2609.25845*, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yizi Zhang, Yanchen Wang, Donato Jiménez-Benetó, Zixuan Wang, Mehdi Azabou, Blake Richards, Renee Tung, Olivier Winter, International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

APPENDIX CONTENTS

A Extended related work
B Data, preprocessing and query definitions
C Calcium forward model
D Architecture, hyper-parameters and baselines
E Proof of Proposition 1
F Additional results: secondary metrics, all signal subsets, training dynamics, per-session analysis, confusion matrices, latent geometry, region importance, transfer, autoregressive exposure bias
G Compute

A EXTENDED RELATED WORK

Behaviour is everywhere in the brain. Brain-wide recordings have established that choice, action and engagement are encoded across most of the forebrain and midbrain (Steinmetz et al., 2019), that spontaneous movements dominate single-trial variance in cortex (Musall et al., 2019; Stringer et al., 2019b), and that visual cortex population responses are high-dimensional (Stringer et al., 2019a). Standardised tasks now allow such findings to be pooled across laboratories (The International Brain Laboratory, 2021), and large surveys map these codes across areas (Siegle et al., 2021a; de Vries et al., 2020). These results motivate a model that asks *many* behavioural questions of *many* regions at once, rather than one decoder per question. They also motivate our region-occlusion readout (Section 4.4), which is a model-based counterpart to area-by-area decoding.

Relating recording modalities. Calcium imaging and electrophysiology differ in temporal resolution, sampling bias and nonlinearity (Wei et al., 2020; Siegle et al., 2021b); spike inference from fluorescence is itself a benchmarked problem (Theis et al., 2016; Friedrich et al., 2017; Pachitariu et al., 2018). The LFP reflects mainly synaptic and transmembrane currents summed over hundreds of micrometres (Mitzdorf, 1985; Buzsáki et al., 2012; Einevoll et al., 2013). Rather than converting one modality into another, brainllax learns sensor-specific stems that map every signal into a common token space and lets downstream attention decide what to combine. Stable latent dynamics across days (Gallego et al., 2020) and the rapid growth of recording capacity (Stevenson & Kording, 2011) suggest such shared spaces should become increasingly useful.

Multimodal binding and generalist models. ImageBind aligns six modalities in one embedding space (Girdhar et al., 2023); Perceiver IO handles arbitrary inputs and outputs with queries (Jaegle et al., 2022); Unified-IO and Gato serialise heterogeneous outputs into tokens (Lu et al., 2023; Reed et al., 2022). Masked modelling (Devlin et al., 2019; He et al., 2022) and patch tokenisation (Dosovitskiy et al., 2021) are the standard pre-training tools that such models build on. brainllax inherits its output interface from Jev teptypesafe2026jev and is closest to Perceiver IO in mechanics, but specialises its output interface to *typed decisions* (softmax, ordinal, Bernoulli) with Null-masking, which is what behavioural neuroscience actually needs.

B DATA, PREPROCESSING AND QUERY DEFINITIONS

Neuropixels. We use the release of Steinmetz et al. (2019) binned at 10 ms (39 sessions, 10 mice, 33,958 units, 14,420 trials of which 10,050 are task trials and 4,370 passive-block trials). The analysis window spans -0.2 to $+0.8$ s relative to stimulus onset. Spike counts are re-binned to 20 ms (50 bins). For every session, units are randomly partitioned into two halves: the first half forms SPK tokens, the second is passed through the calcium model of Section C to form CA* tokens. RGN tokens are computed from all units: for each atlas area present and each 20 ms bin, the fraction of units with at least one spike. LFP tokens are the 100 samples (10 ms) of each LFP channel in the window, standardised per session; the stem additionally receives log power in six bands (1–3, 3–6, 6–12, 12–20, 20–35, 35–50 Hz).

Two-photon calcium. The spontaneous-behaviour recording of Stringer et al. (2019b) provides deconvolved activity for 11,983 neurons at ≈ 3 Hz for 7,018 frames, together with running speed,

pupil area and a face-motion SVD. Neurons are z-scored. Contexts are 10-frame windows with stride 2 (3,504 contexts). Region labels are three imaging-depth terciles. The recording is split into 10 contiguous blocks (blocks 3 and 8 test, block 6 validation, the rest training) and windows within 6 frames of a block boundary are discarded to prevent leakage through overlap.

Splits. Sessions $3, 12, 20, 29, 36$ (mice Forssmann, Lederberg, Moniz, Richards and Tatum) are held out for transfer. Within each remaining session, trials are permuted with a fixed seed and split $70/10/20$.

Table 2: **Decision queries.** ST = Neuropixels, SC = two-photon. Score queries are quantile-binned per session. Null = masked from loss.

Query	Type	K	Source	Definition (Null when)
Choice	choice	3	ST	wheel response: left / no-go / right (passive)
Stimulus side	choice	3	ST	sign of left−right contrast
Behavioural state	choice	3	SC	running > median; else face motion > median; else quiescent
Contrast	score	4	ST	max contrast $\in \{0, .25, .5, 1\}$
Reaction time	score	5	ST	quintile of RT (no-go, passive, invalid)
Wheel speed	score	5	ST	quintile of peak |wheel velocity|, 0–0.8 s
Pupil area	score	5	ST+SC	quintile of mean pupil area in window
Face motion	score	5	ST+SC	quintile of face-motion energy
Running speed	score	5	SC	still (lowest 20%) + quartiles of moving
Reward	bool	2	ST	feedback positive (passive)
Engaged	bool	2	ST	task trial vs. passive-block trial
Previous reward	bool	2	ST	previous trial rewarded (first trial, passive)
Lick	bool	2	ST	any lick 0–0.8 s
Locomotion	bool	2	SC	running speed > median

C CALCIUM FORWARD MODEL

For each unit assigned to CA* we convolve its 10 ms spike train over the full 2.5 s trial with a double-exponential GCaMP6s-like kernel $k(t) \propto (1 - e^{-t/\tau_r})e^{-t/\tau_d}$, τ_r=50 ms, τ_d=0.7 s (Chen et al., 2013), normalised to unit peak; multiply by a per-unit gain drawn from $\log \mathcal{N}(0, 0.4^2)$ to model expression variability; add white Gaussian noise (σ=0.08); and sample at 30 Hz. The 30 frames covering the analysis window are kept. Because the kernel integrates spikes from before the window, CA* tokens carry slow history that SPK tokens do not, and because they come from disjoint units they carry complementary rather than duplicated information.

D ARCHITECTURE, HYPER-PARAMETERS AND BASELINES

Table 3: **Hyper-parameters** (shared by all transformer variants).

width d	192	latents	64
encoder self-attn blocks	4	decoder blocks	2
heads	6	MLP ratio	4
dropout	0.1	channel cap per signal	1,024
optimiser	AdamW	learning rate	5×10^{-4}
weight decay	0.05	warm-up / schedule	300 steps / cosine
batch size	64	epochs	50
modality dropout	0.3	query-set dropout	0.2
gradient clip	1.0	precision	bf16 autocast
parameters	12.7 M	model selection	best validation mean κ

Most parameters ($\approx$ 9.0M) are channel-identity embeddings (46,637 channels $\times$ 192, including the 11,983 imaged neurons); the shared network has $\approx$ 3.7M parameters.

Hyper-parameter selection. We ran two 40-epoch pilots (learning rate 5×10^{-4}, batch 64 vs. 10^{-3}, batch 96); the former reached higher validation κ (0.657 vs. 0.643) and was used for all runs. No other tuning was done.

Per-session linear baseline. Features are time-binned rates (SPK: 5 bins; RGN: 5 bins; CA*: 3 bins; LFP: six log band powers), concatenated over signals, standardised on training trials and reduced by PCA to $\min(256, n_{\text{train}}-1)$ dimensions. The inverse regularisation strength is chosen from $\{0.003, 0.01, 0.03, 0.1, 0.3, 1\}$ by validation accuracy. For the calcium dataset, features are the mean activity of every neuron in the two halves of the window.

Autoregressive variant. The decoder uses a causal mask, and query i adds an embedding of (task$_{i-1}$, answer$_{i-1}$). Training uses the ground-truth answer (teacher forcing), with inapplicable (Null) answers replaced by a uniformly random class (see Section F.2 for why); evaluation feeds back the model's own arg-max answers, one query at a time. Latency is measured without key–value caching for all methods.

E Proof of Proposition 1

brainllax computes $\mathbf{Z}$ once (C_{enc}) and one decoder pass over all N query tokens ($C_{\text{dec}}(N)$); there is no dependency between queries beyond attention inside that pass, so the number of sequential steps is one. In the autoregressive model, answer i is an input to query $i{+}1$, so step $i{+}1$ cannot start before step i finishes: N sequential steps. Without key–value caching, step i re-processes the i-token prefix at cost $C_{\text{dec}}(i) \geq i\, C_{\text{dec}}(1)$ (the per-token feed-forward cost is linear in prefix length), hence $\sum_i C_{\text{dec}}(i) = \Theta(N^2) C_{\text{dec}}(1)$. With caching the total drops to $\Theta(N)$ token-steps but remains N sequential launches. Separate models each pay one encoder and one single-query decoder pass. □

F Additional results

F.1 Secondary metrics and ablations

Table 4 reports type-specific metrics and calibration for all methods. The three-seed brainllax ensemble has the lowest NLL and ECE of all methods and the best Bool AUROC; per-session linear decoders keep the best balanced accuracy on choice queries and the best Spearman ρ on score queries. A single brainllax has lower NLL than the linear decoders but a somewhat higher ECE, and the autoregressive variants are by far the worst calibrated.

Table 4: **Type-specific metrics** averaged over queries of each type (mean over seeds where available). Decoder calls: sequential passes needed to answer N queries.

Method	Choice bal. acc.	Bool AUROC	Score Spearman ρ	mean NLL ↓	mean ECE ↓	decoder calls
Per-session linear	0.724	0.891	0.801	0.731	0.069	N models
Single-task	0.656	0.846	0.750	1.003	0.149	N models
MLP heads	0.691	0.878	0.766	0.707	0.066	1
w/o q-attn	0.697	0.878	0.771	0.722	0.076	1
AR, naive teacher forcing	0.711	0.841	0.770	1.176	0.150	N (sequential)
Autoregressive	0.703	0.871	0.773	0.938	0.113	N (sequential)
brainllax	0.693	0.879	0.770	0.720	0.080	**1**
brainllax, 3-seed ensemble	0.718	0.892	0.792	0.648	0.038	3×1

F.2 A pitfall of autoregressive decision decoding: Null leakage

Our first autoregressive baseline fed the *true* previous answer during training and a dedicated Null token when that answer was inapplicable. This silently leaks information: on passive-block trials choice, reward and reaction time are all Null, so the Null token itself reveals that the animal is disengaged. The model learned this shortcut. At test time, where it must feed back its own (never-Null) predictions, it predicted "engaged" for every trial ($\kappa = 0.000$ on the engagement query; mean $\kappa = 0.593$). Replacing Null answers by a random class during teacher forcing removes the shortcut;

this corrected variant is the "Autoregressive" baseline reported in the main text. Parallel decoding is immune to the problem by construction, because no query ever consumes another query's label.

F.3 All 15 signal subsets

Table 5: **Any-subset inference.** Test κ of the *same* brainllax weights when only the listed signals are provided (mean of three seeds; Neuropixels queries).

Signals	Choice	Stim. side	Contrast	Reward	Engaged	Prev. reward	Reaction time	Wheel speed	Lick	Pupil area	Face motion	Mean
SPK	0.54	0.43	0.59	0.28	0.98	0.33	0.76	0.81	0.70	0.38	0.86	0.604
RGN	0.38	0.28	0.44	0.18	0.85	0.25	0.67	0.75	0.65	0.25	0.79	0.497
LFP	0.16	0.07	0.09	0.09	0.29	0.15	0.35	0.46	0.34	0.03	0.46	0.226
CA	0.51	0.42	0.59	0.26	0.99	0.32	0.75	0.80	0.71	0.43	0.86	0.603
SPK+RGN	0.54	0.44	0.59	0.28	0.98	0.33	0.77	0.81	0.71	0.39	0.86	0.608
SPK+LFP	0.53	0.43	0.58	0.27	0.98	0.33	0.76	0.81	0.71	0.38	0.86	0.604
SPK+CA	0.57	0.46	0.62	0.28	0.99	0.34	0.79	0.82	0.72	0.45	0.88	0.631
RGN+LFP	0.40	0.29	0.43	0.19	0.84	0.26	0.67	0.76	0.64	0.25	0.79	0.500
RGN+CA	0.52	0.42	0.59	0.26	0.99	0.32	0.76	0.80	0.71	0.43	0.87	0.606
LFP+CA	0.51	0.42	0.59	0.26	0.99	0.33	0.76	0.80	0.71	0.43	0.87	0.605
SPK+RGN+LFP	0.54	0.43	0.58	0.27	0.98	0.32	0.77	0.81	0.71	0.39	0.86	0.606
SPK+RGN+CA	0.57	0.46	0.62	0.28	0.99	0.34	0.79	0.82	0.72	0.45	0.88	0.630
SPK+LFP+CA	0.57	0.46	0.62	0.27	0.99	0.35	0.80	0.82	0.73	0.45	0.88	0.631
RGN+LFP+CA	0.52	0.42	0.59	0.26	0.99	0.32	0.77	0.80	0.71	0.44	0.87	0.607
SPK+RGN+LFP+CA	0.57	0.46	0.62	0.28	0.99	0.34	0.80	0.82	0.72	0.45	0.88	0.631

F.4 Training dynamics

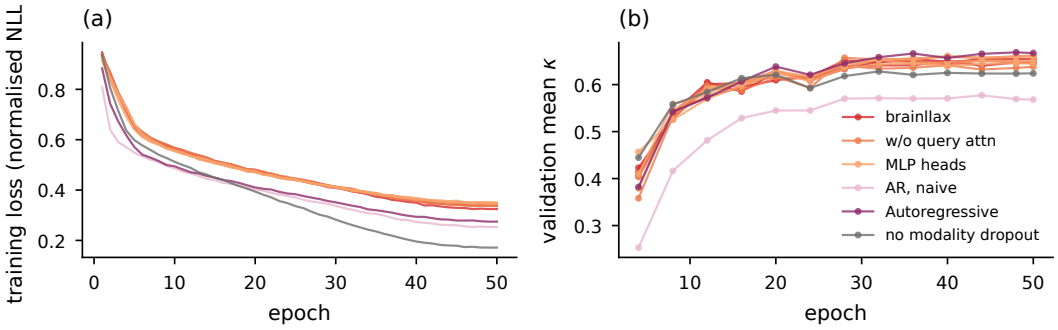

Figure 7: **Training dynamics.** (a) Training loss (type-normalised NLL). (b) Validation mean κ every four epochs. Without modality dropout the model fits the training set best but generalises worst among the full-context models. Autoregressive models reach a low training loss because they see the previous query's true answer, and the naive variant, which can also exploit Null leakage, has by far the lowest validation κ.

F.5 Per-session and per-mouse performance

F.6 Confusion matrices

F.7 Geometry of the global state

F.8 Region importance for all well-sampled regions

F.9 Calibration by query type

G Compute

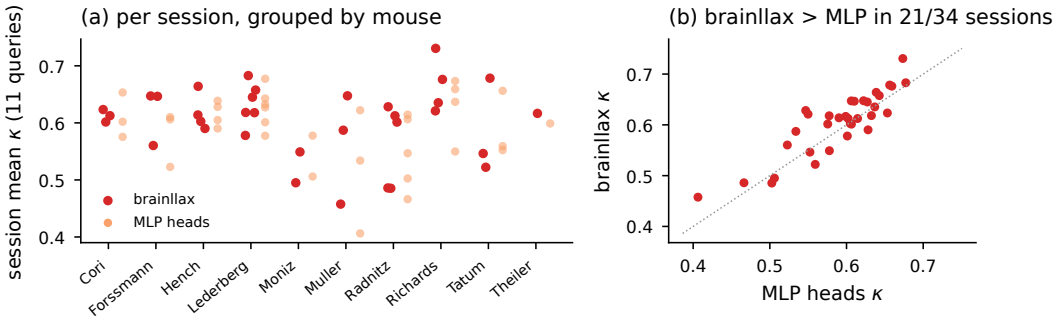

Figure 8: **Session-level results.** (a) Mean κ over the 11 Neuropixels queries for every training session, grouped by mouse. (b) Session-wise comparison of brainllax and the pooled-state MLP-head baseline (seed 0).

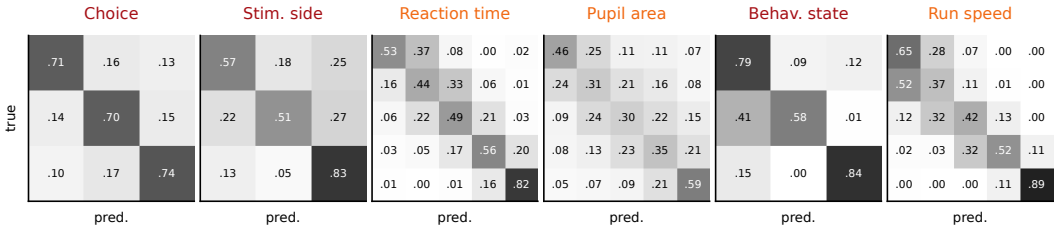

Figure 9: **Row-normalised confusion matrices** of brainllax (seed 0). Ordinal queries (reaction time, pupil, running speed) concentrate errors on adjacent bins, as encouraged by the rank-consistent head.

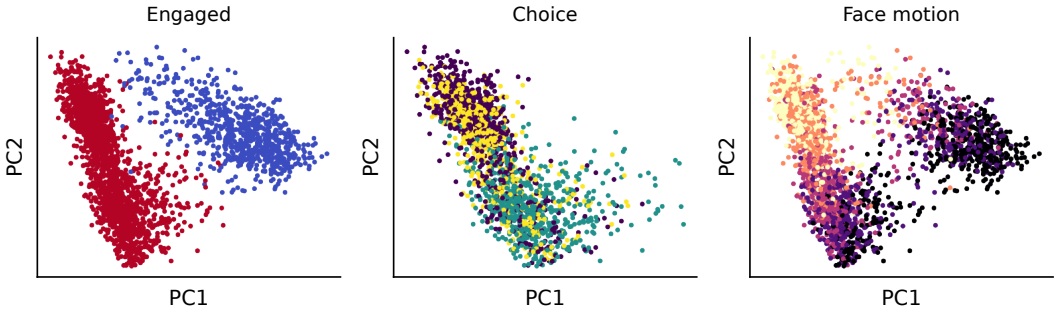

Figure 10: **First two principal components of the mean-pooled global state Z** for Neuropixels test trials, coloured by three query labels. Engagement separates along the leading axis; choice and face motion organise within it.

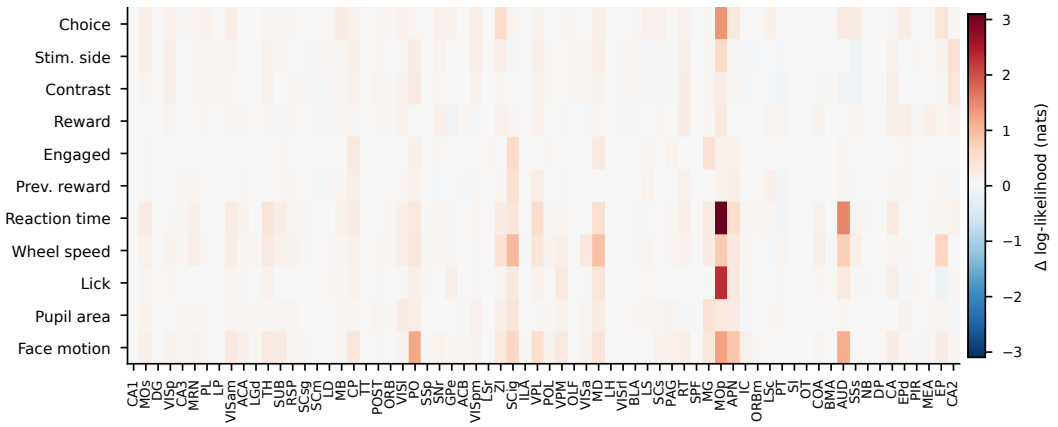

Figure 11: **Region-occlusion importance** (drop in test log-likelihood when every token from a region is removed) for all regions with at least 60 test trials, unnormalised. In absolute terms the largest single effects come from primary motor cortex (MOp) on reaction time, licking and wheel speed, and from auditory cortex (AUD) and the posterior thalamic complex (PO) on reaction time and face motion. The intermediate superior colliculus (SCig) is the top region for engagement and wheel speed.

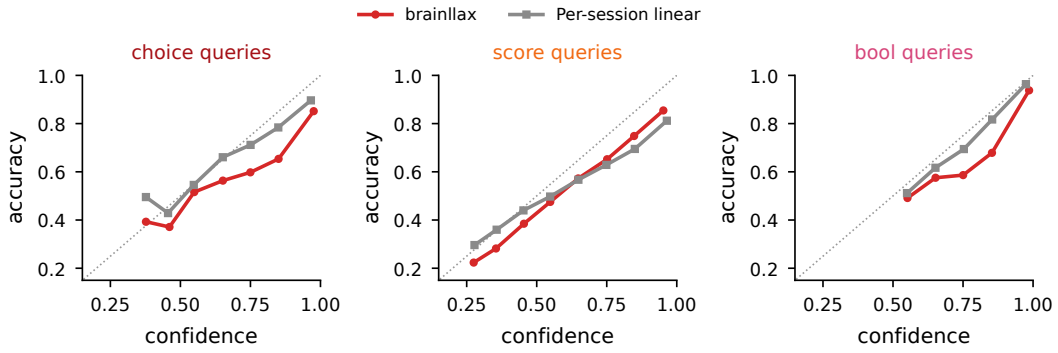

Figure 12: **Reliability diagrams by query type** for brainllax and the per-session linear decoders.

Table 6: **Compute.** Wall-clock per training run on one RTX A6000 shared by three concurrent runs. Summed over all runs, training took 9.23 run-hours. Because up to three runs shared each GPU concurrently, the complete study took 2.4 hours of wall-clock time on the two-GPU server, including downloads, preprocessing ($\approx$ 5 min), all training runs, CPU linear baselines, transfer, the latency benchmark and all analyses.

Run group	# runs	params (M)	train time / run (min)
brainllax (3 seeds)	3	12.65	20.5
w/o query attention	3	12.65	24.5
MLP heads	3	11.50	21.5
Autoregressive (naive + Null-safe)	2	12.67	21.8
No modality dropout	1	12.65	26.3
Single-modality specialists	4	12.65	10.3
Single-task models	14	12.65	16.3
Hyper-parameter pilots	2	12.65	7.9