IMPRINT OF EXPERIENCE: HOW LANGUAGE SHAPES CONCEPT REPRESENTATIONS FROM PRETRAINING TO CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models learn about the world through linguistic experience. Different languages make different aspects of that world explicit, some routinely require speakers to mark when an event occurs or whether a referent is identifiable, while others allow such information to remain implicit in context. This raises a broader question: how does learning experience shape internal concept representations? We connect comparisons across five languages with controlled English training, tracing linguistic experience from pretraining to continual learning. English–Mandarin comparisons show that definiteness and classifier categories remain decodable without corresponding obligatory marking. Yet shared decodability conceals an experiential imprint. In late-layer representations, removing articles lowers full-space definiteness selectivity relative to matched random deletion while raising it in the leading four principal components. The influence of grammatical experience thus depends on which aspects of the representation are examined. Further training reveals a temporal dimension of this imprint. During article-free continuation, an earlier probe becomes less effective, whereas refitting extracts the distinction more successfully from the same updated representations. Together, these findings show that models can represent the same conceptual distinctions while carrying different imprints of linguistic experience. Experience shapes how concept information is organized, and continued learning changes how that information can be recovered.

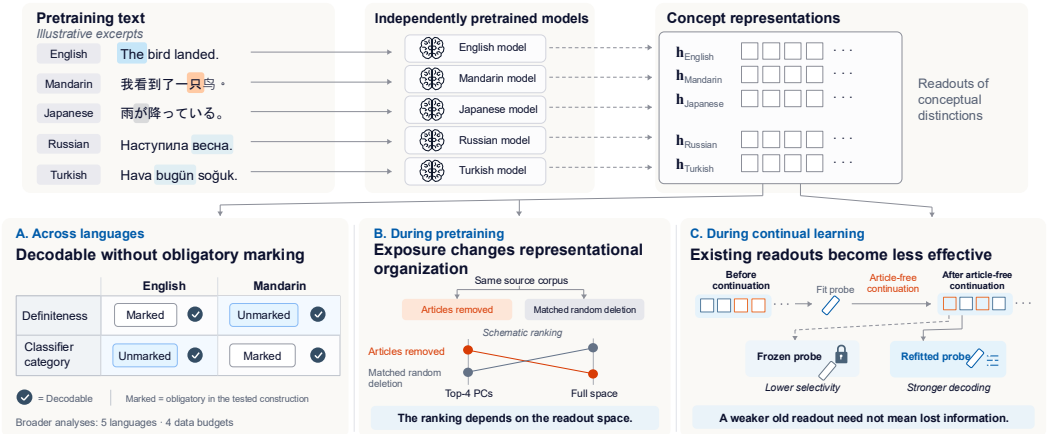

Figure 1: **Shared concepts, different imprints of experience.** **(A)** Conceptual distinctions remain decodable across opposite grammatical marking requirements: English marks definiteness, whereas Mandarin requires numeral classifiers in the tested constructions. **(B)** Article removal reverses the selectivity ranking relative to matched random deletion between leading principal components and full representations. **(C)** Continued training weakens an earlier readout while refitting recovers stronger decoding. Together, these findings trace how linguistic experience shapes the organization of concept information and its accessibility through subsequent learning.

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.

1 INTRODUCTION

Language models learn about the world through linguistic experience. But languages differ in which information their grammar requires speakers to express. English articles distinguish *a bird* from *the bird*, marking whether a referent is presented as identifiable to the listener; Mandarin can express this distinction without an obligatory article system. In the opposite direction, Mandarin requires numeral classifiers in ordinary counting constructions, while English generally permits a numeral to combine directly with a count noun (Dryer & Haspelmath, 2021). These grammatical requirements change the prediction problems a next-token learner encounters during training. An English model must learn when to predict *a* or *the*; a Mandarin model must learn which classifier to predict in a counting expression. How do these recurring choices influence the representations learned from text? This connects a concrete difference between languages to a broader question about language learning: ***how does experience shape internal concept representations?***

One line of research approaches this question by examining the common structure that emerges across models trained on different data. The Platonic Representation Hypothesis proposes convergence toward shared representations of the world (Huh et al., 2024), while embedding translation demonstrates useful correspondences without paired examples (Jha et al., 2026). Independently pretrained monolingual language models can also develop alignable representations without joint multilingual training (Conneau et al., 2020; Zhou et al., 2026). These findings establish commonality across learning experiences, but commonality alone does not explain their influence. Models can represent the same conceptual distinction while differing in how they encode it and how its representation evolves. This matters for both model comparison and knowledge retention: recovering a distinction across models or checkpoints does not reveal whether it is represented in the same way or remains accessible through the same means. We investigate the imprint of linguistic experience within this shared recoverability.

We use grammatical variation to trace this imprint from pretraining to continual learning (Figure 1). Cross-linguistic comparisons of independently pretrained Goldfish models (Chang et al., 2024) span five languages and four data budgets. English and Mandarin provide opposite directions of obligatory marking, allowing us to examine whether the corresponding distinctions remain recoverable across different grammatical systems. To uncover experiential effects beneath this shared recoverability, we manipulate article exposure during English pretraining, holding the architecture and tokenizer fixed and comparing article removal against matched random deletion. We then follow representations through continued language-model training under changes in grammatical exposure. The cross-linguistic comparisons establish shared recoverability, controlled pretraining reveals the influence of grammatical exposure beneath it, and continued training shows how concept representations evolve with further experience.

Shared decodability conceals a measurable imprint of grammatical experience. Definiteness and numeral-classifier category remain decodable in both English and Mandarin models despite their opposite marking requirements. Activation interventions further connect feature-related states to continuation preferences in three feature–language settings. Yet controlled English pretraining reveals a striking reversal. In late-layer representations, removing articles lowers full-space definiteness selectivity relative to matched random deletion, while raising it in the leading four principal components. Both directions are also observed in task accuracy, linking the crossover to the prediction of definiteness itself. The same grammatical experience therefore has different effects on how readily the distinction can be recovered from different aspects of the representation. This *decodability crossover* reveals an influence that a single decoding score would obscure. Measurement diagnostics support these comparisons by identifying lexical shortcuts and label-dependent cross-language rankings.

Further learning shows why the imprint of experience must also be understood over time. Models receiving equal exposure to article-preserving and article-deleted corpora reach different final levels of decodability when their order is reversed, demonstrating that the sequence of experience matters alongside its composition. During article-free continuation, a previously fitted probe becomes less effective, while a newly fitted probe extracts the distinction more successfully from the same updated hidden states. These results distinguish retaining concept information from preserving an established way of extracting it. Together with the pretraining experiments, they show that shared conceptual knowledge can carry different histories of experience, expressed in its internal organization and in how it remains accessible as learning continues.

2 Related Work

Linguistic experience and concept representations. Training-data manipulations reveal how linguistic experience shapes learning, through shuffled word order (Sinha et al., 2021), transformed grammatical rules (Kallini et al., 2024), and controlled bilingual exposure (Dufter & Schütze, 2020). Goldfish enables complementary comparisons across independently pretrained monolingual models (Chang et al., 2024). Interpreting learned representations also requires distinguishing information from its extraction (Belinkov, 2022): control tasks calibrate probe performance (Hewitt & Liang, 2019), while information-theoretic and description-length approaches characterize information and extraction difficulty (Pimentel et al., 2020; Voita & Titov, 2020). Targeted corpus ablations ask what models still learn from indirect evidence once a construction is removed (Misra & Mahowald, 2024; Patil et al., 2024). We connect these perspectives through controlled grammatical exposure, asking how removing a grammatical channel reshapes the representation of what is still learned.

Shared representations across languages and models. Multilingual transfer (Pires et al., 2019; Wu & Dredze, 2019) and post-hoc monolingual alignment (Conneau et al., 2020) reveal structure shared across languages. Recent work develops this perspective through representational convergence (Huh et al., 2024), unpaired embedding translation (Jha et al., 2026), shared semantics across languages and modalities (Wu et al., 2025), and functional transfer between independently pretrained models (Zhou et al., 2026). Causal alignment and subspace-patching studies further examine how representational correspondences relate to behavior (Geiger et al., 2024; Makelov et al., 2024). We investigate the imprint of experience beneath this commonality: how a distinction that remains recoverable across linguistic systems is represented.

Representational change during continual learning. Continual learning studies knowledge acquisition and retention (Kirkpatrick et al., 2017), while continued pretraining adapts language models to new domains and tasks (Gururangan et al., 2020). Newly fitted classifiers distinguish representation forgetting from deterioration in existing predictions (Davari et al., 2022); related work preserves linear separability as features change (Gu et al., 2023) and examines the generality of pretrained representations through continued learning (Luo et al., 2023). In neuroscience, *representational drift* describes population codes that change while behavior stays stable, so that fixed decoders degrade unless refitted (Driscoll et al., 2017; Rule et al., 2019). We bridge retention to the history of grammatical exposure and find an analogous drift in language models: a concept can remain recoverable as an earlier means of extracting it becomes less effective.

3 Tracing Experience Through Concept Representations

Recovering the same distinction across models leaves open how their different experiences shape its representation. We connect three views of this relationship: decodability across grammatical systems, organization under controlled exposure, and readout stability as learning continues. Together, these comparisons trace what linguistic experience leaves behind beyond the presence of a recoverable distinction.

3.1 Three Views of Linguistic Experience

Across languages. Independently pretrained Goldfish models (Chang et al., 2024) allow us to compare grammatical systems without joint multilingual training. We study English, Mandarin, Russian, Turkish, and Japanese at budget tiers of 5, 10, 100, and 1,000 MB, retaining each model's original corpus and tokenizer. English and Mandarin reverse the marking direction for definiteness and numeral-classifier category (Table 1). Temporal distinctions extend coverage to all five languages, while Turkish evidentiality provides a within-language case. English–Mandarin temporal comparisons additionally match label definitions and cue conditions.

During pretraining. We train 125M-parameter English models from scratch on variants of a shared corpus drawn from the Goldfish English source dataset, holding the architecture and existing English tokenizer fixed. We vary article retention and compare complete article removal against random deletion of non-article function words, matched in deletion count and sentence-initial count. This contrast distinguishes removing articles from removing a comparable amount of other grammatical text. Training budgets and corpus construction appear in Appendix A.4.

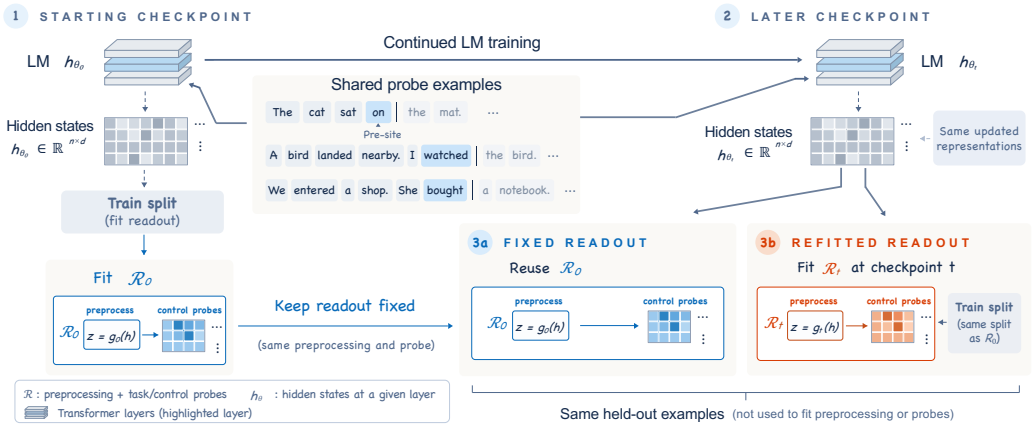

Figure 2: **Separating readout stability from recoverability.** Shared examples yield pre-site representations at the starting and later checkpoints. The fixed path reuses the original preprocessing and classifiers $\mathcal{R}_0$; the refitted path estimates $\mathcal{R}_t$ on updated training representations. Both evaluate the same later hidden states on identical held-out examples, at the same layer and regularization. Text examples are illustrative.

Table 1: **Language–feature coverage.** The shaded block highlights opposite marking directions. Controlled pretraining and continuation manipulate English article exposure.

Distinction	English	Mandarin	Russian	Turkish	Japanese
Definiteness	**M**	**U**	—	—	—
Numeral-classifier category	**U**	**M**	—	—	—
Past/non-past	M	U	M	M	M
Evidentiality	—	—	—	M†	—

M: marking required in the tested construction; **U**: no corresponding requirement, although lexical or contextual cues may remain; —: outside the primary analysis. †Within-Turkish analysis only. Classifier labels distinguish *ge* from other classifiers, with English noun labels assigned through a bilingual lexicon.

During continual learning. Two sequential designs examine subsequent change. The *order* experiment reverses article-preserving and article-deleted training while holding total exposure to each corpus fixed. The *common-fork* experiment branches five models after 32,650 steps of article-preserving pretraining into article-deleted or random-deletion continuation. Each pair shares starting weights, optimizer and scheduler states, and random-number-generator states, followed by matched update budgets and learning-rate schedules.

3.2 Reading Out Concept Representations

Decodability. A model with parameters θ is trained on an *experience* $E = (D_1, \ldots, D_m)$, an ordered sequence of corpora. For an example x with label y, $h_\theta(x) \in \mathbb{R}^d$ is the *pre-site* hidden state at the last token preceding the target site, excluding the target from the causal prefix. A readout $\mathcal{R} = (g, c, \tilde{c})$ consists of preprocessing g (standardization, followed by projection onto the leading k principal components when applicable) and L_2-regularized linear task and control classifiers ($C = 1$), fitted on class-balanced examples. Following Hewitt & Liang (2019), the control classifier $\tilde{c}$ is trained on labels permuted by a class-frequency-preserving π, and selectivity subtracts its held-out accuracy from task accuracy:

$$S(\mathcal{R}; \theta) = \mathrm{Acc}_{\mathrm{test}}\big(c \circ g \circ h_\theta,\ y\big) - \mathrm{Acc}_{\mathrm{test}}\big(\tilde{c} \circ g \circ h_\theta,\ \pi(y)\big). \tag{1}$$

Writing $\hat{\mathcal{R}}_k(\theta)$ for the readout fitted to θ with k retained components ($k = d$ for the full state), the *readout profile* is $S_\theta(k) = S(\hat{\mathcal{R}}_k(\theta); \theta)$. Cross-language, from-scratch, and order experiments use stratified five-fold evaluation; common-fork experiments use shared 16k/4k/20k training/validation/test splits. Probe cells require at least 100 examples per class.

Organization across readout spaces. We compare standardized full states with projections onto their leading k principal components, fitting standardization and PCA only on probe-training data; standardization keeps a few high-magnitude dimensions from dominating the leading components

(Timkey & Van Schijndel, 2021). Cross-validated profiles use the layer maximizing full-space cross-validation selectivity, shared across all k. Common-fork analyses select the layer by validation selectivity at the fork and hold it fixed thereafter. For experiences E and E' that differ in one grammatical channel, the *imprint* at readout dimension k and its interaction across readout spaces are

$$\Delta(k) = \mathbb{E}\big[S_{\theta_E}(k)\big] - \mathbb{E}\big[S_{\theta_{E'}}(k)\big], \qquad I = \Delta(4) - \Delta(d), \tag{2}$$

with expectations over pretraining seeds. Shared decodability constrains only $\Delta(d)$; a *decodability crossover* occurs when $\Delta(k)$ and $\Delta(d)$ take opposite signs for some $k < d$, an imprint that no single readout space registers in full.

Readout stability. Continued training moves the fork model θ_0 to θ_t. We evaluate each later representation using the fork readout $\mathcal{R}_0 = \hat{\mathcal{R}}_d(\theta_0)$ and a newly fitted readout $\mathcal{R}_t = \hat{\mathcal{R}}_d(\theta_t)$ (Figure 2):

$$S_t^{\text{fixed}} = S\big(\mathcal{R}_0; \theta_t\big), \qquad S_t^{\text{refit}} = S\big(\mathcal{R}_t; \theta_t\big). \tag{3}$$

The fixed readout preserves the entire pipeline fitted at the fork. Refitting re-estimates that pipeline on updated hidden states using the same training split, layer, and regularization, without updating the language model. Both readouts use identical held-out examples; the primary continuation analysis uses full states. Since both equal S_0 at the fork, the change in the established readout splits exactly into a change in recoverability and a readout gap:

$$S_t^{\text{fixed}} - S_0 = \big(S_t^{\text{refit}} - S_0\big) - G_t, \qquad G_t = S_t^{\text{refit}} - S_t^{\text{fixed}}. \tag{4}$$

Equation 4 is an accounting identity whose value is that both terms are separately measurable. A *readout–recoverability dissociation* is a decline in the established readout carried by G_t while recoverability is maintained, which a frozen readout alone cannot distinguish from forgetting.

Validation and uncertainty. Embedding-layer and carrier-held-out diagnostics assess lexical shortcuts, while activation interventions examine functional relevance. Controlled contrasts use pretraining seeds as the replication unit, with Welch intervals for independent conditions and paired t intervals for within-seed comparisons. We report 95% intervals; longitudinal intervals are pointwise. Item-bootstrap intervals condition on the fitted readout and control accuracy. Appendix A.3 details the evaluation procedures.

4 Results

4.1 Different Languages, Shared Conceptual Distinctions

The tested distinctions remain recoverable across opposing grammatical requirements. Mandarin models encode definiteness without an obligatory article system, while English models encode classifier category without obligatory numeral classifiers. Together, these comparisons establish shared decodability across two directions of linguistic variation.

Decodability in both marking directions. At 1,000 MB, definiteness selectivity reaches 0.268 in Mandarin and 0.263 in English (Figure 3). Mandarin selectivity rises from 0.075 at the embedding layer to 0.268 at the best contextual layer, so the distinction is built through context. Classifier category supplies the reverse test: selectivity is 0.188 in English and 0.189 in Mandarin for the shared *ge*/other inventory, exceeding their respective embedding baselines of 0.140 and 0.115. Obligatory marking is therefore unnecessary for positive readouts in either comparison.

Measurement distinguishes context from lexical shortcuts. A noun-inclusive readout reaches 0.810 accuracy on English classifier category at the embedding layer, exposing how noun identity can support decoding before contextual processing. Likewise, English clause-tense and temporal-carrier labels agree on only 59% of carrier-bearing items; changing the definition reverses the ordering relative to Mandarin under matched cue conditions. Pre-site readouts and matched definitions make the comparison interpretable (Appendix A.3).

Across languages and budgets. Definiteness and classifier readouts generally strengthen between 5 and 1,000 MB, with depth profiles separating embedding-level recoverability from contextual contributions. Temporal and evidential analyses extend the comparison to Russian, Turkish, and Japanese (Appendix A.2).

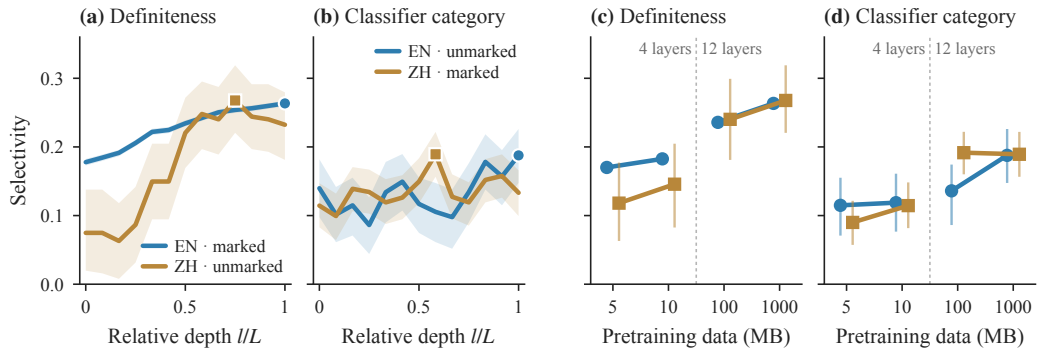

Figure 3: **Decodable in both marking directions.** (a,b) Pre-site selectivity across relative depth at 1,000 MB, including embedding-layer baselines at $l/L = 0$. (c,d) Best-layer selectivity across training budgets. Balanced sample counts at 1,000 MB are 40,000/254 for English/Mandarin definiteness and 522/856 for classifier category. Architecture varies across budget tiers; evaluation details and uncertainty appear in Appendix A.2.

Table 2: **Feature-related states influence continuations.** DAS at $k = 16$ on the probe-selected layer. ΔIIA is the gain over the unpatched baseline; Random is IIA under dimension-matched random-subspace intervention.

Setting	n_{test}	IIA	ΔIIA	Random	Support
English / definiteness	795	.736	+.330	.418	Met
Mandarin / definiteness[†]	840	.767	+.525	.231	Met
English / temporal reference	900	.753	+.290	.459	Met
Mandarin / temporal reference[‡]	192	.938	+.229	.724	Below floor

[†]Expanded UD-plus-Wikipedia intervention set; primary probing uses UD items. [‡]Insufficient pair support. Test counts refer to held-out ordered pairs. Random-subspace controls use a separate evaluation subset. Details appear in Appendix A.6.

From decodable states to continuation preferences. Distributed alignment search (DAS; Geiger et al., 2024) identifies subspaces in which source-to-base activation interchange shifts continuation preferences toward the source label. At $k = 16$, interventions improve interchange intervention accuracy (IIA) in three supported feature–language settings (Table 2), connecting recoverable distinctions to model predictions.

Shared without joint training. As a multilingual reference, XGLM-564M (Lin et al., 2022) differs from the monolingual models by less than 0.025 selectivity in 13 of 15 cells on matched evaluation subsets. Within XGLM, Mandarin exceeds English by 0.097 for definiteness and 0.067 for classifier category. The same-sign advantage across opposite marking directions resists a simple account based on obligatory marking alone. The monolingual results independently establish that joint multilingual training is unnecessary for the observed recoverability.

Shared decodability establishes what these models can represent. We next ask what it conceals about the experience through which their representations were formed.

4.2 Shared Decodability Conceals an Imprint of Grammatical Exposure

Grammatical exposure leaves an imprint that a single decoding score can miss. Removing articles makes definiteness harder to decode from the full late-layer representation, yet easier to decode from its leading principal components. The direction of the apparent benefit therefore reverses with the readout space.

The same experience has opposite effects across readout spaces. Relative to matched random deletion, article removal lowers full-space definiteness selectivity while raising top-4 selectivity: $\Delta(d) = -0.0131$ and $\Delta(4) = +0.0205$, an interaction of $I = +0.0336$ (Eq. 2; Table 3). The same reversal holds against article-preserving training (Figure 4). Both directions also appear in

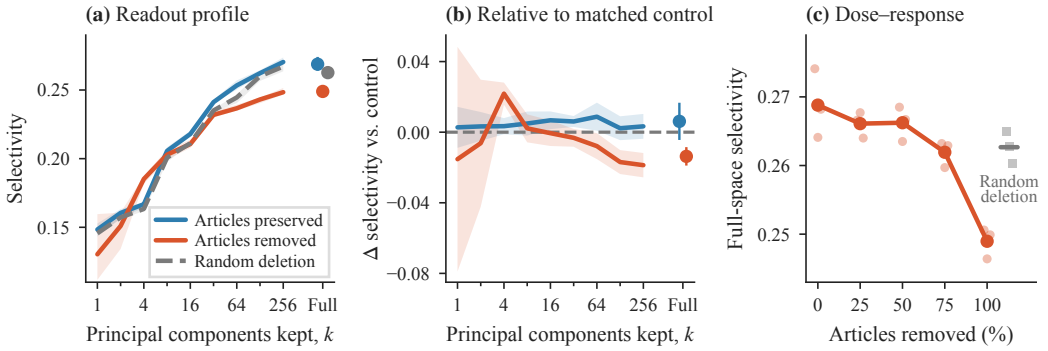

Figure 4: **Article exposure reshapes the readout profile.** (a) Selectivity across retained principal components; curves use three seeds per condition, with five-seed preserved/removed means at $k = 4$, $k = 16$, and Full. (b) Contrasts against matched random deletion show opposite effects in full space and leading components. (c) Full-space selectivity across article-deletion rates, with three seeds per rate and a random-deletion reference.

Table 3: **Grammatical exposure reverses the readout ranking.** Late-layer definiteness selectivity on 40k shared examples. Condition values are mean ± SD across pretraining seeds; contrasts report point estimates with Welch 95% CIs beneath. Contrasts report $\Delta(d)$, $\Delta(4)$, and the interaction I of Eq. 2.

Pretraining corpus (seeds)	Full space	Top-4 PCs
Articles preserved (5)	0.2674 ± 0.0047	0.1680 ± 0.0027
Random deletion (3)	0.2627 ± 0.0024	0.1634 ± 0.0019
Articles removed (5)	0.2496 ± 0.0029	0.1839 ± 0.0043

Condition contrast	Full space, $\Delta(d)$	Top-4 PCs, $\Delta(4)$	Interaction I
Removed − random	-0.0131 $[-0.0178, -0.0083]$	$+0.0205$ $[+0.0150, +0.0260]$	$+0.0336$ $[+0.0252, +0.0419]$
Removed − preserved	-0.0178 $[-0.0236, -0.0119]$	$+0.0159$ $[+0.0105, +0.0213]$	$+0.0336$ $[+0.0244, +0.0429]$
Task accuracy: removed − random			
	-0.0147 $[-0.0180, -0.0115]$	$+0.0226$ $[+0.0176, +0.0276]$	—

task accuracy, linking the crossover to prediction of definiteness itself rather than to control-task calibration. Each effect exceeds the seed-to-seed standard deviation four- to tenfold, and the finding is a reversal of sign rather than a difference in magnitude.

A decodability crossover. Article exposure thus produces a *decodability crossover* (Section 3.2): against both references, $\Delta(4)$ and $\Delta(d)$ take opposite signs. Article exposure improves decoding from the full late-layer state, while article removal improves decoding along its dominant variance directions. Grammatical experience therefore changes the relationship between a recoverable distinction and the representation's dominant structure, which a single score would register as either a benefit or a cost.

Grammatical exposure connects representations to prediction. At article-prediction sites in English, definiteness is part of the next-token decision itself: *a* or *the*. Obligatory marking makes the distinction relevant to what the model must prepare to express, providing a computational parallel to *thinking for speaking* (Slobin, 1996). Our output-direction analysis gives this connection a concrete representational basis: article-preserving and random-deletion models align their definiteness readouts more strongly with the article-output direction, which alone supports decoding close to the full-space probe. Removing this direction nevertheless leaves most of the full-space condition gap recoverable by refitted probes, and the crossover persists (Appendix A.9). Grammatical experience thus shapes the connection between conceptual distinctions and their explicit expression, with representational effects extending beyond the single output direction examined here.

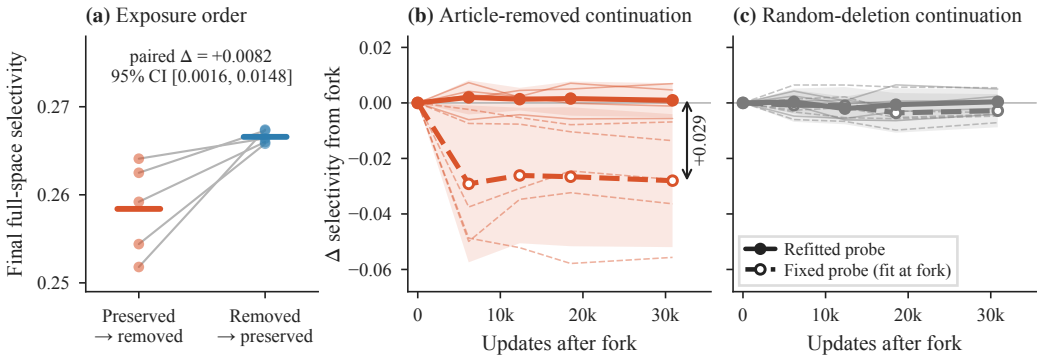

Figure 5: **Experience after pretraining.** (a) Reversing exposure order changes final selectivity despite equal exposure to both corpora. (b,c) Changes from the common fork under article-deleted and random-deletion continuation. Solid curves use refitted readouts; dashed curves retain the fork readout. Bands are pointwise paired t 95% intervals over five seeds. The endpoint gap compares refitted and fixed readouts on the same representations.

The imprint varies nonlinearly with exposure. Intermediate deletion rates reveal how the full-space effect develops. Deleting 25% or 50% of articles leaves mean selectivity within 0.003 of the article-preserving condition; deleting 75% lowers it by 0.0069, and complete removal by 0.0198 (Figure 4c). The imprint thus emerges as article evidence disappears rather than in proportion to the evidence removed.

Experience thus leaves an imprint beneath shared decodability. We next ask how that imprint responds when experience changes during continued learning.

4.3 Continued Learning Changes How Concept Information Is Read

Retaining a conceptual distinction does not guarantee that an established decoder will continue to recover it: sequential training reveals a dependence on exposure order and a separation between an earlier readout's effectiveness and the information recoverable afterward.

Exposure order shapes the endpoint. Models trained on the same two corpora reach higher full-space selectivity when article-preserving text comes last rather than first: 0.2666 versus 0.2584, a paired difference of 0.0082 with 95% CI $[0.0016, 0.0148]$. Under the shared training schedule, equal exposure totals therefore produce different endpoints when their order is reversed. The composition of experience alone does not specify the representation reached.

Established readouts become less effective. During article-free continuation, Eq. 4 separates the two terms: the fork readout's change of -0.0280 equals a recoverability change of only $+0.0010$ minus a readout gap of $G_t = 0.0290$ (Table 4). Task accuracy shows the same separation. Under random deletion, mean changes in both readouts remain close to zero.

A readout–recoverability dissociation. Refitting changes the measurement pipeline while leaving the language model untouched, so its advantage is decoding that the updated representation supports but the earlier readout no longer reaches. Together with the pretraining crossover, this shows that tracing experience requires examining both concept information and the means used to extract it.

4.4 Discussion

Shared information can carry different histories. Representational convergence emphasizes common structure across systems (Huh et al., 2024). Our results identify an experiential dimension within this commonality. Models can recover the same conceptual distinction while grammatical exposure changes which aspects of their representations support stronger decoding. The crossover makes this relationship concrete: full-space decoding favors article exposure, whereas decoding from leading components favors article removal. Agreement about what is recoverable can therefore coexist with different conclusions about how experience supports its recovery.

Table 4: **Experience changes the endpoint and the effectiveness of an established readout.** Full-space definiteness selectivity under continued training. Values are paired mean differences at the final checkpoint with 95% CIs beneath, using five seeds per experiment; columns follow Eq. 4. Fixed readouts retain the fork's preprocessing and classifiers; refitted readouts re-estimate both on updated representations.

Continuation	Fixed readout $S_t^{\text{fixed}} - S_0$	Refitted readout $S_t^{\text{refit}} - S_0$	Readout gap G_t
Articles removed	-0.0280 $[-0.0520, -0.0041]$	$+0.0010$ $[-0.0052, +0.0071]$	$+0.0290$ $[+0.0074, +0.0505]$
Random deletion	-0.0028 $[-0.0088, +0.0032]$	$+0.0004$ $[-0.0040, +0.0049]$	$+0.0032$ $[-0.0041, +0.0105]$
Article − random	-0.0252 $[-0.0506, +0.0002]$	—	—
Task accuracy under article-deleted continuation			
	-0.0304 $[-0.0541, -0.0067]$	$+0.0008$ $[-0.0032, +0.0048]$	$+0.0312$ $[+0.0083, +0.0542]$

Exposure order	Final selectivity	Paired difference
Preserved → removed	0.2584	
Removed → preserved	0.2666	$+0.0082$ $[+0.0016, +0.0148]$

Article − random compares within-seed changes between continuation branches. The exposure-order difference compares removed→preserved against preserved→removed under equal exposure totals. Dashes indicate contrasts not tabulated.

Retention includes the stability of access. Continual-learning research distinguishes deterioration in an existing classifier from representation forgetting (Davari et al., 2022); Eq. 4 measures both under controlled changes in linguistic experience. A readout carries assumptions established at the moment it was fitted; subsequent learning can make those assumptions less effective while leaving the distinction recoverable. Pairing an established readout with a newly fitted reference therefore separates two practically different outcomes: the information a current representation supports and the effectiveness of an earlier means of extracting it. This distinction matters whenever an existing decoder or monitoring tool is carried across model updates.

Training history is part of the representation's description. The order experiment shows that aggregate exposure does not fully characterize the learning experience: the same corpora, encountered in different orders, produce different endpoints. Together with the crossover and readout dissociation, this motivates describing representations through both their training history and their decoding profile. Grammatical marking provides a concrete experimental variable for connecting properties of experience to what models encode and how that information remains accessible.

5 Conclusion and Limitations

Different linguistic experiences can support the same conceptual distinctions while leaving different internal expressions of them. Cross-language comparisons establish shared decodability; controlled pretraining reveals an imprint whose apparent direction reverses across readout spaces; and continued training separates recoverable information from the effectiveness of an earlier decoder. Together, these findings show why understanding the imprint of experience requires tracing both concept representations and the changing means of accessing them.

Limitations. Our controlled experiments examine English article exposure in 125M-parameter models with three to five seeds per condition. Extending these interventions across languages, grammatical systems, and model scales would establish their broader reach. The analyses characterize linear recoverability and readout stability; behavioral transfer and adaptation provide complementary targets for future investigation. Cross-language comparisons retain the released models' corpora and tokenizers, while PCA-based profiles jointly vary retained directions and dimensionality. The appendices document annotation, sampling, intervention support, robustness analyses, and behavioral-task outcomes.

AI USE STATEMENT

Generative AI tools were used solely to assist with writing the experimental code. All other aspects of this work were completed by the authors without generative AI assistance. The authors reviewed and tested the AI-assisted code to verify its correctness. The authors take full responsibility for the final content of this work, including the AI-assisted code.

ETHICS STATEMENT

This work studies linguistic representations in language models using existing models and text corpora. The experiments concern computational systems rather than human cognition, and the findings should not be interpreted as evidence of differences in cognitive ability among speakers or of the superiority of any language. Cross-language comparisons may reflect biases and uneven coverage in the source corpora, as well as choices in tokenization, annotation, and evaluation. Our conclusions are therefore restricted to the tested models, grammatical distinctions, and experimental settings. We recognize that existing corpora and pretrained models may contain social biases, and our analyses do not establish the fairness or safety of these models for downstream deployment.

REPRODUCIBILITY STATEMENT

Section 3 describes the experimental designs, readout procedures, and uncertainty estimation. Appendix A.4 details the training budgets and corpus construction for controlled pretraining, while Appendix A.3 documents the evaluation procedures and measurement diagnostics. Appendix A.2 provides details of the cross-language experiments, and Appendix A.6 describes the activation interventions and multilingual reference comparisons. Together, these details specify the training and evaluation settings needed to reproduce the reported analyses.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. Goldfish: Monolingual language models for 350 languages. *arXiv preprint arXiv:2408.10441*, 2024.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 6022–6034, 2020.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16691–16700. IEEE, 2022.

Laura N Driscoll, Noah L Pettit, Matthias Minderer, Selmaan N Chettih, and Christopher D Harvey. Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell*, 170(5):986–999, 2017.

Matthew Dryer and Martin Haspelmath. The world atlas of language structures online. *Max Planck Institute for Evolutionary Anthropology*, 2021.

Philipp Dufter and Hinrich Schütze. Identifying elements essential for bert's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4423–4437, 2020.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

Qiao Gu, Dongsub Shim, and Florian Shkurti. Preserving linear separability in continual learning by backward feature projection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24286–24295. IEEE, 2023.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8342–8360, 2020.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John Morris. Harnessing the universal geometry of embeddings. *Advances in Neural Information Processing Systems*, 38:45963–45987, 2026.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14691–14714, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 9019–9052, 2022.

Yun Luo, Zhen Yang, Xuefeng Bai, Fandong Meng, Jie Zhou, and Yue Zhang. Investigating forgetting in pre-trained representations through continual learning. *arXiv preprint arXiv:2305.05968*, 2023.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Kanishka Misra and Kyle Mahowald. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 913–929, 2024.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajic, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 4034–4043, 2020.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. Filtered corpus training (fict) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615, 2024.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, 2020.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4996–5001, 2019.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pp. 101–108, 2020.

Michael E Rule, Timothy O'Leary, and Christopher D Harvey. Causes and consequences of representational drift. *Current opinion in neurobiology*, 58:141–147, 2019.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 2888–2913, 2021.

Dan I. Slobin. From "thought and language" to "thinking for speaking". In John J. Gumperz and Stephen C. Levinson (eds.), *Rethinking Linguistic Relativity*, pp. 70–96. Cambridge University Press, 1996.

William Timkey and Marten Van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 833–844, 2019.

Zhaofeng Wu, Xinyan Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *International Conference on Learning Representations*, volume 2025, pp. 53705–53723, 2025.

Ej Zhou, Suchir Salhan, Catherine Arnett, and Anna Korhonen. Cross-lingual alignment without joint training: Do monolingual language models converge on universal representations? *arXiv preprint arXiv:2608.27115*, 2026.

A Appendix

A.1 Data and Annotation

Every probe item carries a label defined identically in the languages it compares, derived from treebank annotation or from a rule-based labeler applied to parsed text. This section specifies the sources, the label and site for each distinction, and how the labels were verified.

Sources. Items come from Universal Dependencies release 2.15 (Nivre et al., 2020), pooling all splits: English EWT, Mandarin GSDSimp, Russian SynTagRus, Turkish BOUN and IMST, and Japanese GSD. Stanza-parsed Wikipedia sentences (Qi et al., 2020) enter a cell only when the parser and labeler reproduce gold UD labels with item-level $F_1 \geq 0.90$ and label agreement ≥ 0.95; this admits English definiteness (F_1 0.991), English tense (0.949), and Mandarin definiteness (0.918). Items are not translation-parallel across languages.

Definiteness. The label is the overt marker of the upcoming noun phrase, and the pre-site position is the token immediately before it. English uses the determiner lemma: *the* for definite, *a/an* for indefinite. Mandarin uses 这/那/该 with a classifier for definite and 一 with a classifier for indefinite, excluding series selectors such as 最后 and month names. Because Mandarin marking is optional, 425 of the 16,328 core noun phrases in GSDSimp qualify. An expanded UD-plus-Wikipedia Mandarin set (19,174 items) supplies intervention pairs; it is more separable at the embedding layer (selectivity 0.221 against 0.075), so probing uses the UD set.

Numeral-classifier category. A Mandarin noun-to-classifier lexicon is derived from GSDSimp `clf` arcs, excluding measure, time, and event units, and retains 140 nouns. English nouns (109) receive a class through the MUSE English–Mandarin dictionary (Conneau et al., 2017) when all of their translations agree. The Mandarin site is a classifier slot opened by a numeral or determiner; the English site is a lexicon noun with a numeral or determiner on its left. Neither pre-site position has seen the noun. Only *ge* (个) reaches 100 items in both languages, so the tested distinction is *ge* versus other classifiers.

Past/non-past. Two label definitions are used. *Clause tense* (English, Russian, Turkish) is the `Tense` value of the clause's finite element: the root if it is finite and tensed, otherwise the leftmost finite auxiliary or copula; `Fut` and English *will*/*shall* count as non-past. Japanese items are past when the auxiliary た attaches to the predicate. *Temporal carriers* are lexicon entries (temporal adverbials, plus aspect markers for Mandarin), each with a class and a carrier group (day, week, month, year, span, elapsed, general, or aspect). Mandarin labels are the carrier's class, and the site is the first verb after the carrier or a predicate marked by a native speaker. English carriers must attach to the labeled predicate within its clause. Items whose carrier directly precedes the site are removed, and carrier-to-site distances are binned as 1–2, 3–5, and 6+ tokens, with subsets that contain no intervening finite verb. On the English pool, the two definitions agree on 59.0% of carrier-bearing items.

Evidentiality. The Turkish label is the UD `Evident` value (firsthand or non-firsthand) of the root predicate, which is also the site; missing values are not read as firsthand.

Sampling. Pools larger than 40,000 items are balanced by label and subsampled to 40,000, stratified by source and carrier group. Sites beyond the 512-token context are dropped, and the majority class is down-sampled to 1:1. A cell requires at least 100 items per class (Table 5).

Table 5: **Probe cells at 1,000 MB.** Pool: labeled items before the 40,000 cap. Probed: items with a locatable site. Balanced: items after class balancing. Check: native-speaker agreement on a stratified sample.

Distinction	Lang.	Source	Pool	Probed	Balanced	Check
Definiteness	EN	EWT + Wiki	1,612,273	40,000	40,000	—
	ZH	GSDSimp	425	425	254	.970 / .960
	ZH, expanded[a]	GSDSimp + Wiki	19,174	19,174	10,550	—
Classifier	EN	EWT + lexicon	1,423	1,414	522	—
	ZH	GSDSimp	978	978	856	—
Past/non-past	EN	EWT + Wiki	616,180	39,979	39,974	—
	ZH	GSDSimp + Wiki[b]	1,046	1,046	962	lexicon, sites
	RU	SynTagRus	56,043	40,000	40,000	—
	TR	BOUN + IMST	7,964	7,914	7,126	.995
	JA	GSD	5,624	5,622	4,332	.995
Evidentiality	TR	BOUN + IMST	2,477	2,445	1,104	1.000

[a]Used for interventions only. [b]227 UD items plus 819 Wikipedia sentences with native-speaker-marked sites.

Verification. English labels and Russian tense come directly from treebank annotation or from a parser validated against it. Native speakers checked stratified samples of the remaining cells. Mandarin definiteness agreed on 194 of 200 items (0.970); all disagreements involved 一 in series or name expressions, which the exclusion rule above removes, and the final rule agrees on 0.960 of 100 fresh items. Turkish and Japanese tense agree on 0.995 and Turkish evidentiality on 1.000 of 200-item samples. For Mandarin carriers, annotators retained 27 of 30 lexicon patterns (4 under syntactic conditions) and marked the carrier's predicate in Wikipedia sentences, yielding 819 items.

Carrier-held-out gap. To test whether a tense readout depends on specific carrier words, carrier groups are split into halves A and B by a deterministic greedy assignment; the gap is cross-validated accuracy within A minus the accuracy on B of a probe fitted on A. At 1,000 MB the gap is 0.085 for English, -0.113 for Russian, and 0.145 for Mandarin. Other cells lack a half with 100 items.

A.2 Cross-Language Results

The main text reports 1,000 MB endpoints. Here we give every cell at every budget beside its embedding-layer value, which separates lexical from contextual recoverability, and show that the English–Mandarin tense ordering follows the label definition.

Models and readouts. Each cell uses one released Goldfish model. Models at 5 and 10 MB have 4 layers of width 512; models at 100 and 1,000 MB have 12 layers of width 768. Budget comparisons are therefore made within this suite and share no scale with the controlled models. All cells use the pre-site position, the probe of Section 3.2, and stratified five-fold cross-validation. The grid covers definiteness and classifier category in English and Mandarin, past/non-past in all five languages, and Turkish evidentiality, each at four budgets (Tables 6 and 7, Figure 6).

Table 6: **Definiteness and classifier category across budgets.** Best-layer pre-site selectivity with the selected layer in parentheses; the $l = 0$ row gives the embedding-layer value of the same cell. M/U as in Table 1; n: balanced items.

Cell		n	Row	5 MB	10 MB	100 MB	1,000 MB
Definiteness							
English	M	40,000	best	.1701 (4/4)	.1826 (3/4)	.2358 (12/12)	.2634 (12/12)
			$l = 0$	.1672	.1748	.1750	.1778
Mandarin	U	254	best	.1181 (4/4)	.1457 (0/4)	.2402 (11/12)	.2677 (9/12)
			$l = 0$	.0236	.1457	.1339	.0748
Numeral-classifier category (ge vs. other)							
English	U	522	best	.1149 (1/4)	.1188 (4/4)	.1360 (12/12)	.1877 (12/12)
			$l = 0$	.0747	.1054	.1207	.1398
Mandarin	M	856	best	.0900 (0/4)	.1145 (4/4)	.1916 (8/12)	.1893 (7/12)
			$l = 0$	.0900	.1063	.0935	.1145

Definiteness and classifier category. At 1,000 MB, the Mandarin definiteness readout rises from 0.075 at the embedding layer to 0.268 (item-bootstrap 95% intervals $[0.020, 0.138]$ and $[0.221, 0.319]$), so the distinction is built through context in a language that does not require it. English classifier category reaches 0.188 against an embedding value of 0.140; the two intervals ($[0.148, 0.226]$ and $[0.098, 0.182]$) overlap, so its contextual gain rests on the point estimates. The English and Mandarin classifier endpoints differ by 0.0016.

Table 7: **Past/non-past and Turkish evidentiality across budgets.** Format as in Table 6. *gap*: carrier-held-out gap at the best layer, where both halves reach 100 items.

Cell		n	Row	5 MB	10 MB	100 MB	1,000 MB
Past/non-past							
English	M	39,974	best	.2688 (4/4)	.2831 (4/4)	.3274 (12/12)	.3397 (12/12)
			$l = 0$	.2307	.2352	.2535	.2572
Mandarin	U	962	best	.2973 (4/4)	.2661 (1/4)	.3950 (10/12)	.4761 (10/12)
			$l = 0$	.1445	.1580	.1674	.1632
			gap	.2133	.2027	.1435	.1445
Russian	M	40,000	best	.1145 (4/4)	.1333 (4/4)	.1744 (10/12)	.1961 (10/12)
			$l = 0$	.0914	.0991	.0986	.1042
			gap	.0835	.0199	$-.0462$	$-.1131$
Turkish	M	7,126	best	.0729 (2/4)	.0868 (3/4)	.1368 (12/12)	.1591 (10/12)
			$l = 0$	.0644	.0614	.0794	.0779
Japanese	M	4,332	best	.1127 (2/4)	.1166 (4/4)	.1766 (12/12)	.2198 (12/12)
			$l = 0$	.0900	.0600	.0669	.0669
Evidentiality (firsthand vs. non-firsthand)							
Turkish	M	1,104	best	.0761 (0/4)	.1078 (3/4)	.0996 (5/12)	.1594 (12/12)
			$l = 0$	.0761	.0082	.0543	.0525

Past/non-past depends on the label definition. Whole-cell tense values are within-language descriptions: every Mandarin label comes from a temporal carrier, whereas the other languages label clause morphology. Matching carrier distance and intervening verbs isolates the comparison (Ta-

ble 8). Under clause-tense labels, Mandarin exceeds English by up to 0.19 in the bins without intervening verbs. Under carrier-class labels, the definition Mandarin uses, the two languages agree within 0.033 in all five computable bins. The cross-language ordering therefore follows how the label is defined (Figure 7b).

Table 8: **Carrier-matched past/non-past at 1,000 MB.** Pre-site selectivity at each subset's best layer (balanced n). Bins are carrier-to-site distances in tokens; "no int." excludes intervening finite verbs. Last column: Mandarin minus English under carrier-class labels.

Bin	English, clause tense	English, carrier class	Mandarin	Mand. − Eng.
1–2	.3547 (468)	.4957 (932)	.5288 (416)	+.033
3–5	.2896 (404)	.4414 (614)	.4551 (312)	+.014
6+	.4106 (358)	.4041 (532)	.3889 (234)	−.015
1–2, no int.	.3715 (358)	.5166 (722)	.4850 (400)	−.032
3–5, no int.	.2248 (218)	.4383 (324)	.4118 (272)	−.027

A.3 Measurement Protocol

Two choices determine what a readout reveals: the token position it reads and the definition of its label (Figure 7). This section fixes both, together with the probe, the control, and the uncertainty estimates shared by all experiments.

Extraction. Hidden states are extracted in fp32 in evaluation mode. Layer 0 is the embedding output, and depth is reported as l/L. Each Goldfish model and XGLM-564M is probed with its own tokenizer; the controlled models share the unchanged Goldfish English tokenizer. Sequences are truncated at 512 tokens.

Pre-site position. The pre-site token is the last non-special token whose character span ends at or before the first character of the target site; a causal model computes its hidden state without access to the site. For definiteness the site is the upcoming marker, and for classifier category it is the classifier slot (Mandarin) or the pre-noun slot (English). Two pooled positions serve as references: the mean over content tokens up to the pre-site token, and the mean over all content tokens, which includes the target and is the target-inclusive readout of Figure 7a.

Probe and control. Every readout applies standardization, optional PCA to k components without whitening, and L_2-regularized logistic regression with $C = 1$ (lbfgs, at most 2,000 iterations). The control fits the same pipeline, with the same C and k, to one fixed class-frequency-preserving permutation of the balanced labels (Hewitt & Liang, 2019). Truncating the control together with the task probe matters, because a rank-k probe has less capacity to fit permuted labels. Selectivity is task minus control accuracy (Eq. 1).

Cross-validated readouts. Cross-language, from-scratch, and exposure-order readouts use stratified five-fold cross-validation, refitting the scaler and PCA basis in each training fold. Accuracy is computed over all out-of-fold predictions. The reported layer maximizes full-space selectivity over all layers, including layer 0, and the k-curve is read at that layer.

Common-fork readouts. A fixed readout must be frozen on one training set, so the continuation experiment splits the 40,000 balanced items once into 16,000 training, 4,000 validation, and 20,000 test items. The layer with the highest validation selectivity at the fork is fixed for all later checkpoints of that seed. $\mathcal{R}_0$ is fitted there and applied unchanged, while $\mathcal{R}_t$ refits the same pipeline on the checkpoint's own training split (Eq. 3); both are scored on the same test items. Single-split and cross-validated selectivities are compared only within their own regime.

Items. The from-scratch and continuation experiments share one set of 40,000 English definiteness items (20,000 per class; 39,603 from Wikipedia and 397 from EWT). The exposure-order trajectory uses a separate 15,000-item subsample, and the XGLM comparison uses matched subsamples capped at 15,000 items.

Uncertainty. For a single released model, the interval is a percentile bootstrap over items (1,000 resamples) of task correctness at the reported layer, shifted by the observed control accuracy. Contrasts between independently trained conditions use Welch intervals with the pretraining seed as the

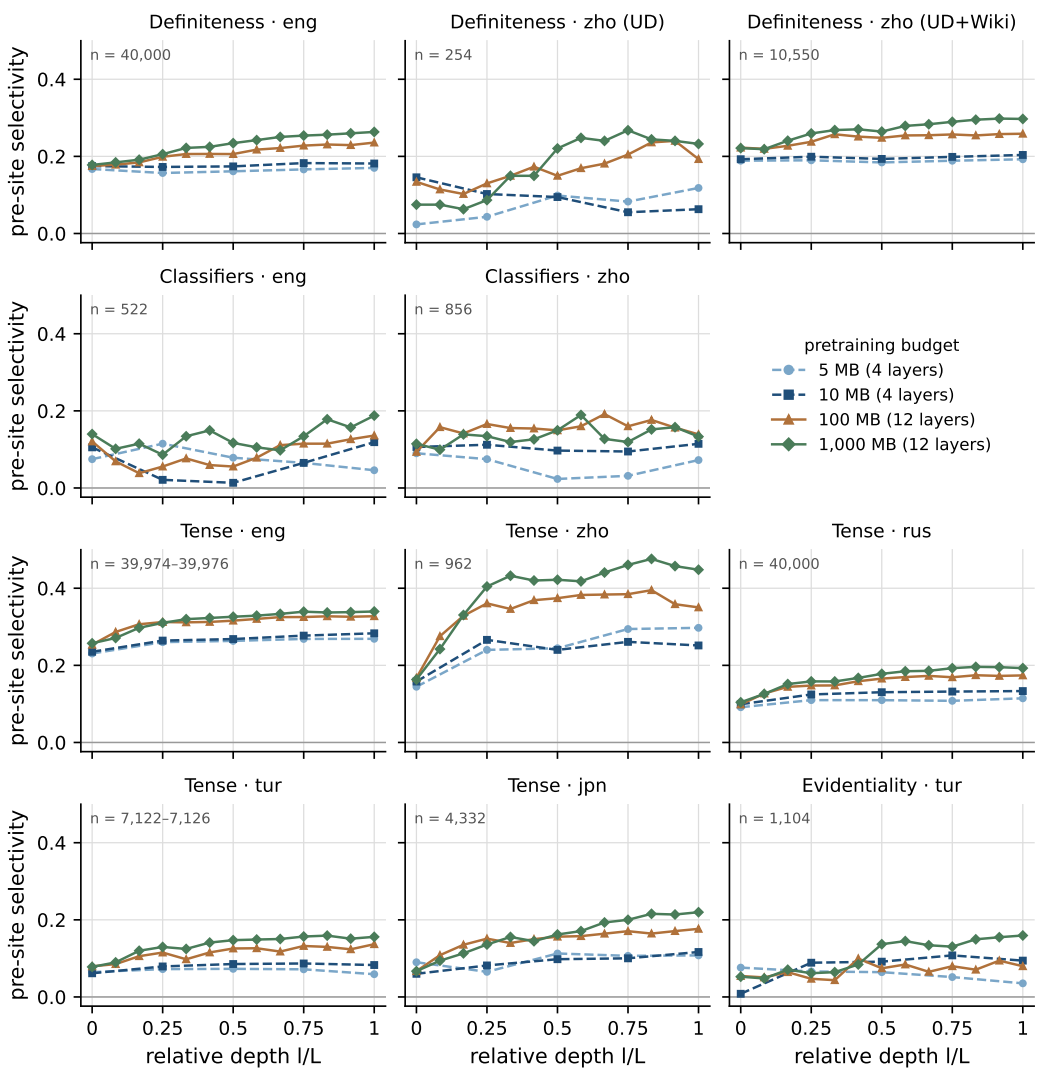

Figure 6: **Per-layer profiles of every cross-language cell.** Pre-site selectivity against relative depth l/L; $l/L = 0$ is the embedding output. Dashed lines: 4-layer models (5 and 10 MB); solid lines: 12-layer models (100 and 1,000 MB). All panels share one vertical axis; n is the number of balanced items.

unit. Within-seed contrasts (exposure orders, common-fork changes, and refitted-minus-fixed gaps) use paired t intervals, computed pointwise per checkpoint. All intervals are 95%.

A.4 Controlled Pretraining

The controlled corpora are renderings of a single source text that differ only through rule-based deletions, so that any difference between models traces to the words removed. This section specifies the corpora, the training recipe, and the per-seed results behind Section 4.2.

Source corpus. The source is drawn from the file Goldfish uses for English: the first 1,500 MiB, permuted once by line and cut where packing reaches 417,924 examples of 512 tokens, the training-example count of Goldfish `eng_latn_1000mb`. The article-preserving rendering packs to 213,977,088 tokens per epoch. The next 16,000 packed examples form a held-out evaluation set. All conditions keep the same line order.

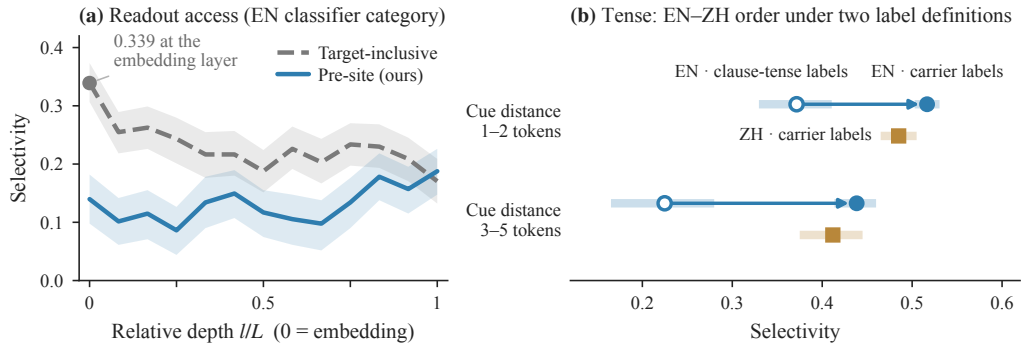

Figure 7: **Readout position and label definition change the diagnosis.** (a) English classifier category across depth for a target-inclusive mean-pooled readout and for the pre-site readout; the target-inclusive readout already reaches 0.810 accuracy at the embedding layer ($n = 522$), so it reports lexical identity. (b) English–Mandarin past/non-past without intervening finite verbs under English clause-tense and carrier-class labels: the two definitions reverse the ordering relative to Mandarin. Balanced n: English 358/722 (distance 1–2) and 218/324 (3–5) for clause/carrier labels; Mandarin 400 and 272. All models are 1,000 MB Goldfish.

Table 9: **Readout settings and selected layers** ($L = 12$ throughout; counts of models or checkpoints in parentheses).

Experiment	Balanced items	Evaluation	Selected l/L	Intervals
Cross-language, 1,000 MB[a]	40,000 / 254 / 522 / 856	5-fold CV	1.0 / 0.75 / 1.0 / 0.58	Item bootstrap
Controlled pretraining and dose	40,000 (shared)	5-fold CV	1.0 (20 of 22); 0.92 (2)	Welch t
Exposure order, endpoints	40,000 (shared)	5-fold CV	1.0 (9 of 10); 0.92 (1)	Paired t
Exposure order, trajectory	15,000	5-fold CV	1.0 (39 of 60); 0.67–0.92 (21)	Means
Common fork	40,000 (16k/4k/20k)	Validation at fork, then fixed	0.75, 1.0, 1.0, 0.83, 1.0	Paired t

[a]English and Mandarin definiteness, English and Mandarin classifier category.

Article deletion. Lines are annotated with Stanza. An article site is a word tagged DET with lemma *the*, *a*, or *an*; demonstratives and possessives are never sites. A deletion removes the article and its following whitespace; nothing is inserted, replaced, or re-capitalized, and no language model generates or rewrites any training text. The source contains 15,909,923 sites, 1,256,249 of them sentence-initial. Rate r deletes $\mathrm{round}(rN)$ sites taken as a prefix of one seeded permutation, so the deleted sets are nested ($25\% \subset 50\% \subset 75\% \subset 100\%$).

Random-deletion control. The control deletes words tagged ADP, AUX, or SCONJ (most often *of*, *in*, *is*, *for*, and *to*), excluding any article form. It deletes exactly as many words as the article-removed corpus (15,909,923), stratified to match the sentence-initial count (1,256,249). The two corpora therefore remove comparable amounts of grammatical text while differing in which words are removed.

Architecture and optimization. Every model copies the configuration of `eng_latn_1000mb`: a GPT-2 decoder with 12 layers, 12 heads, width 768, context 512, vocabulary 51,200, and tied embeddings (124,770,816 parameters), trained from random initialization. All conditions share Goldfish's SentencePiece English tokenizer, unchanged. Optimization follows Goldfish: AdamW (learning rate 10^{-4}, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-6}$, weight decay 0.01), linear warm-up over 10% of steps and linear decay to zero, gradient clipping at 1.0, and batches of 64 sequences. Each run

covers 10 epochs of its own corpus in bf16 autocast with fp32 master weights, taking 1.3–1.5 hours on one NVIDIA B200. For a given seed, initial weights are bit-identical across conditions.

Token budgets. Because the epoch count is fixed, deletion also reduces the number of processed tokens (Table 10). Complete removal and random deletion differ by at most 0.02%, so their contrast is budget-matched; the preserved condition and the intermediate rates change budget together with article exposure.

Table 10: **Controlled-pretraining conditions.** Deleted words and sentence-initial deletions; packed 512-token examples per epoch; optimizer steps; processed tokens relative to the article-preserving condition. Ranges span seeds.

Condition	Deleted	Sentence-initial	Examples/epoch	Steps	Tokens	Seeds
Article-preserving (0%)	0	0	417,924	65,300	0	0–4
25% deleted	3,977,481	313,510–314,161	412,097–412,179	64,390–64,402	−1.4%	0–2
50% deleted	7,954,962	627,654–628,568	406,384–406,413	63,497–63,502	−2.8%	0–2
75% deleted	11,932,442	941,648–942,188	400,644–400,664	62,600–62,603	−4.1%	0–2
Article-removed (100%)	15,909,923	1,256,249	394,908	61,704	−5.5%	0–4
Random deletion	15,909,923	1,256,249	394,915–394,995	61,705–61,717	−5.5%	0–2

Consistency with Goldfish. Article-preserving models reproduce the released model they copy: at 1,000 MB their pre-site definiteness selectivity is 0.2688 against 0.2634 for `eng_latn_1000mb`, and a 100 MB fp32 replicate reaches held-out loss 4.752 against Goldfish's 4.830. Relative to fp32, bf16 changes loss by 0.0002 and probe accuracy by at most 0.006.

Endpoint readout profiles. The selected layer is layer 12 for every model except article-removed seeds 0 and 4 (layer 11), and all readout spaces of a model use the same layer. The crossover holds seed by seed: article-removed models fall below article-preserving models in the full space and rise above them at $k = 4$ in all five seeds (Table 11). Against random deletion, the difference is -0.0131 $[-0.0178, -0.0083]$ in the full space and $+0.0205$ $[+0.0150, +0.0260]$ at $k = 4$, and passes through zero at $k = 16$ (-0.0002 $[-0.0030, +0.0025]$).

Table 11: **Endpoint selectivity by seed** (pre-site, 40,000 items, selected layer). k: leading principal components; Full: untruncated state.

	Article-preserving			Article-removed			Random deletion		
Seed	k=4	k=16	Full	k=4	k=16	Full	k=4	k=16	Full
0	0.1684	0.2165	0.2741	0.1874	0.2137	0.2506	0.1655	0.2113	0.2650
1	0.1644	0.2170	0.2641	0.1865	0.2104	0.2464	0.1628	0.2114	0.2627
2	0.1675	0.2203	0.2682	0.1817	0.2078	0.2499	0.1618	0.2109	0.2603
3	0.1720	0.2207	0.2619	0.1864	0.2107	0.2474	—	—	—
4	0.1676	0.2210	0.2685	0.1773	0.2123	0.2537	—	—	—
Mean	0.1680	0.2191	0.2674	0.1839	0.2110	0.2496	0.1634	0.2112	0.2627
SD	0.0027	0.0022	0.0047	0.0043	0.0022	0.0029	0.0019	0.0003	0.0024

Dose–response. Full-space selectivity declines by -0.0175 per unit deletion rate ($[-0.0205, -0.0145]$, cluster-robust over seeds). Deleting 25% or 50% of articles leaves selectivity at the preserved level, whereas 75% and complete removal lower it by 0.0069 and 0.0198 (Table 12). Held-out loss on undeleted text rises with the deletion rate, confirming that the manipulation reaches the models.

A.5 Exposure Order and Common-Fork Continuation

Both continuation designs hold total exposure fixed and vary only when, or whether, articles are present. This section specifies the designs and gives the per-seed results behind Section 4.3.

Exposure order. Each run trains from scratch in two stages of five passes, one over the article-preserving stream (32,650 steps) and one over the article-removed stream (30,852 steps), in either order. Both orders run 63,502 steps under one linearly decaying schedule, so the corpus seen second

Table 12: **Dose–response** (pre-site, full space, selected layer, three seeds). Δ: contrast with 0% and cluster-robust 95% interval. Loss: mean held-out loss on undeleted text.

Condition	Seed 0	Seed 1	Seed 2	Mean	Δ vs. 0% [95% CI]	Loss
0% (preserved)	0.2741	0.2641	0.2682	0.2688	—	4.061
25%	0.2666	0.2677	0.2640	0.2661	-0.0027 [-0.0103, $+0.0049$]	4.075
50%	0.2685	0.2635	0.2667	0.2662	-0.0026 [-0.0061, $+0.0010$]	4.101
75%	0.2632	0.2597	0.2629	0.2619	-0.0069 [-0.0116, -0.0022]	4.148
100% (removed)	0.2506	0.2464	0.2499	0.2490	-0.0198 [-0.0241, -0.0156]	4.709
Random deletion	0.2650	0.2627	0.2603	0.2627	-0.0061	4.159

is trained at lower learning rates, and runs sharing a seed share their initialization. Ending with article-preserving text yields higher final selectivity in every seed (Table 13). Within runs, adding articles in the second stage raises selectivity by 0.0238 $[0.0153, 0.0323]$, whereas removing them changes it by -0.0034 $[-0.0141, +0.0074]$.

Table 13: **Exposure order by seed.** Full-space selectivity at the endpoints (40,000 items) and change over the second stage (15,000-item trajectory). P: article-preserving; R: article-removed. Intervals are paired t over seeds.

	Endpoints			Second-stage change	
Seed	P→R	R→P	R→P − P→R	P→R	R→P
0	0.2641	0.2664	$+0.0023$	-0.0044	$+0.0138$
1	0.2544	0.2658	$+0.0114$	-0.0050	$+0.0236$
2	0.2592	0.2660	$+0.0068$	-0.0104	$+0.0236$
3	0.2518	0.2674	$+0.0156$	-0.0085	$+0.0331$
4	0.2625	0.2673	$+0.0048$	$+0.0115$	$+0.0250$
Mean	0.2584	0.2666	$+0.0082$	-0.0034	$+0.0238$
95% CI	—	—	$[0.0016, 0.0148]$	$[-0.0141, +0.0074]$	$[+0.0153, +0.0323]$

Common fork. Five seeds train on the article-preserving stream for 32,650 steps of a 63,502-step schedule and save their full training state: weights, optimizer and scheduler states, random-number-generator states, and data cursor. Each branch restores this state and trains 30,852 further steps on either the article-removed stream or a random-deletion stream matched in deletion count and sentence-initial deletions. Both branches process 1,974,528 sequences with identical per-step learning rates, and the restored weights match the saved state exactly. Readouts are taken at the fork and at steps 38,820, 44,990, 51,161, and 63,502.

Continuation results. Under article-free continuation the fixed readout declines in every seed (Table 14); the decline is already -0.0292 $[-0.0574, -0.0009]$ at the first checkpoint and persists through the endpoint, while the refitted readout stays at its fork level. Held-out loss on undeleted text rises from 4.165–4.170 at the fork to 4.639–4.657 after article removal and falls to 4.132–4.136 after random deletion. The difference between the two branches' fixed-readout changes is negative at every checkpoint (-0.0283, -0.0251, -0.0230, -0.0252), with pointwise intervals that include zero.

A.6 Interventions and Multilingual Reference

Interventions test whether the states that probes read influence the model's predictions, and the XGLM comparison tests whether joint multilingual training changes decodability. Both use the 1,000 MB Goldfish models.

Continuation tasks. Each prefix is followed by two candidate continuations that differ only in the distinction: *the* versus *a/an* for English definiteness, a demonstrative versus 一 before the same classifier for Mandarin definiteness, and past versus non-past predicate forms for English temporal reference. In these settings the prefix ends at the pre-site position. For Mandarin temporal reference, which has no verbal inflection, the candidates are opposite-class carriers from the same group and

Table 14: **Common-fork continuation by seed.** Full-space selectivity on the fixed 20,000-item test split. ΔS: change from the fork to step 63,502. Means and intervals are paired t over seeds.

Seed	100	101	102	103	104	Mean	95% CI
Reference layer	9	12	12	10	12		
Selectivity at fork	0.23635	0.25110	0.25050	0.25145	0.25205		
Article-removed branch							
Fixed ΔS	-0.00685	-0.02760	-0.03635	-0.01360	-0.05565	-0.0280	$[-0.0520, -0.0041]$
Refit ΔS	-0.00115	-0.00010	$+0.00465$	$+0.00695$	-0.00560	$+0.0010$	$[-0.0052, +0.0071]$
G_t	$+0.00570$	$+0.02750$	$+0.04100$	$+0.02055$	$+0.05005$	$+0.0290$	$[+0.0074, +0.0505]$
Random-deletion branch							
Fixed ΔS	-0.00710	$+0.00545$	-0.00350	-0.00415	-0.00475	-0.0028	$[-0.0088, +0.0032]$
Refit ΔS	$+0.00240$	$+0.00075$	$+0.00500$	-0.00200	-0.00410	$+0.0004$	$[-0.0040, +0.0049]$
G_t	$+0.00950$	-0.00470	$+0.00850$	$+0.00215$	$+0.00065$	$+0.0032$	$[-0.0041, +0.0105]$

the prefix ends before the carrier. Unpatched models prefer the correct candidate on 64–93% of items.

DAS and control. A learned orthogonal rotation (pyvene `LowRankRotatedSpaceIntervention`) swaps the first $k = 16$ rotated coordinates from the source's last prefix token into the base, at the block with the highest pre-site selectivity. With the language model frozen, the rotation is trained with Adam (10^{-3}, 16 pairs per batch, 20 epochs) on $\mathrm{softplus}(-m)$, where m is the post-patch logit difference between the source's correct and incorrect candidates. IIA is the fraction of held-out pairs with $m > 0$, and ΔIIA subtracts the unpatched rate on the same pairs. Opposite-label items are paired in both directions, and 30% of ordered pairs are held out. The control is the same intervention, untrained, at the same k. A setting has sufficient support with at least 1,400 training and 500 test pairs.

Table 15: **DAS at $k = 16$ on the probe-selected layer.** Task pref.: unpatched preference for the correct candidate; Unpatched: IIA without intervention; IIA with 95% Wilson interval.

Setting	l/L	Train / test	Task pref.	Unpatched	IIA [95% CI]	ΔIIA	Random	Support
English / def.	1.00	1,857 / 795	.7962	.406	.7358 [.7041, .7653]	$+.3295$	.4176	Met
Mandarin / def.[a]	0.92	1,962 / 840	.8518	.242	.7667 [.7369, .7940]	$+.5250$	.2310	Met
English / temporal	1.00	2,100 / 900	.6444	.463	.7533 [.7241, .7803]	$+.2900$	.4589	Met
Mandarin / temporal[b]	0.83	450 / 192	.9265	.708	.9375 [.8939, .9639]	$+.2292$	.7240	Below floor

[a]Expanded UD-plus-Wikipedia set. [b]Carrier-position variant.

Robustness. ΔIIA remains positive at $k \in \{1, 4\}$ on the selected layer and at $k = 16$ on layers $l/L \in \{0.5, 0.75\}$, with a minimum of $+0.209$ in the supported settings. On the UD-only Mandarin definiteness set (254 pairs), ΔIIA is $+0.395$. In every setting the lower Wilson bound of IIA exceeds both the unpatched and the random-subspace rates.

XGLM reference. XGLM-564M (Lin et al., 2022) has 24 layers and was trained jointly on many languages. Each cell probes XGLM and Goldfish on the identical label-balanced subsample, with pre-site positions recomputed under XGLM's tokenizer (Table 16). XGLM stays within 0.025 of Goldfish in 13 of 15 matched cells (the ten shown and five givenness cells), and exceeds it in the two small Mandarin cells. Within XGLM, Mandarin exceeds English by 0.097 for definiteness and 0.067 for classifier category, whereas Goldfish scores the two languages within 0.018 and 0.002 on the same items. Because XGLM also differs in size, data, and tokenization, these differences characterize the model rather than joint training alone.

A.7 Robustness of the Crossover and the Readout Gap

The crossover is read at each model's best layer with one regularization strength, and the readout gap compares two complete readout pipelines. We test whether the crossover depends on regular-

Table 16: **Pre-site selectivity on matched subsets.** L0: embedding-layer selectivity; l/L: best layer; Δ: XGLM − Goldfish.

			XGLM-564M			Goldfish 1,000 MB			
Cell	M/U	n_{bal}	Sel.	L0	l/L	Sel.	L0	l/L	Δ
English / definiteness	M	15,000	.2651	.1669	0.75	.2497	.1669	0.83	+.0154
Mandarin / definiteness	U	254	.3622	.0984	0.79	.2677	.0748	0.75	+.0945
English / classifier	U	522	.2126	.1284	0.83	.1877	.1398	1.00	+.0249
Mandarin / classifier	M	856	.2792	.1799	1.00	.1893	.1145	0.58	+.0899
English / past (clause tense)	M	14,988	.3328	.2270	1.00	.3301	.2455	1.00	+.0027
Mandarin / past (carrier)	U	962	.4688	.1320	0.50	.4761	.1632	0.83	−.0073
Russian / past	M	15,000	.1921	.0795	0.79	.1854	.1029	1.00	+.0067
Turkish / past	M	7,156	.1622	.0520	0.88	.1591	.0779	0.83	+.0031
Japanese / past	M	4,334	.2024	.0985	0.92	.2198	.0669	1.00	−.0174
Turkish / evidentiality	M	1,104	.1540	.0752	0.75	.1594	.0525	1.00	−.0054

ization or depth, whether the leading components behave differently from arbitrary low-dimensional readouts, and which accuracy term carries each effect.

Setup. The nine controlled models of seeds 0–2 are re-read at a fixed middle layer (hidden state 6 of 12, $l/L = 0.5$) with $C \in \{0.1, 1, 10\}$, in the full space and the top four principal components, using the same 40,000 items, folds, and control as the main analysis. The re-extracted states reproduce the stored layer-6 values of all nine models to four decimals.

Table 17: **The crossover across regularization and depth.** Article-removed minus random-deletion selectivity (Welch 95% CI, pretraining seed as the unit; n per condition in parentheses). The first row is the main analysis.

Readout setting	Full space	Top-4 PCs	Interaction I
Best layer, seeds 0–4 (5/5/3)	-0.0131 $[-0.0178, -0.0083]$	$+0.0205$ $[+0.0150, +0.0260]$	$+0.0336$ $[+0.0252, +0.0419]$
Best layer, seeds 0–2 (3/3/3)	-0.0137 $[-0.0189, -0.0085]$	$+0.0218$ $[+0.0156, +0.0281]$	$+0.0355$ $[+0.0255, +0.0455]$
Layer 6, $C = 0.1$ (3/3/3)	-0.0057 $[-0.0107, -0.0007]$	$+0.0177$ $[-0.0178, +0.0532]$	$+0.0234$ $[-0.0158, +0.0626]$
Layer 6, $C = 1$ (3/3/3)	-0.0060 $[-0.0113, -0.0008]$	$+0.0176$ $[-0.0180, +0.0531]$	$+0.0236$ $[-0.0164, +0.0636]$
Layer 6, $C = 10$ (3/3/3)	-0.0055 $[-0.0105, -0.0006]$	$+0.0177$ $[-0.0180, +0.0534]$	$+0.0232$ $[-0.0169, +0.0634]$

Table 18: **Task and control terms of the crossover.** Article-removed minus random-deletion differences (Welch 95% CI).

Setting	Space	Selectivity	Task accuracy	Control accuracy
Best layer	full	-0.0131 $[-0.0178, -0.0083]$	-0.0147 $[-0.0180, -0.0115]$	-0.0016 $[-0.0049, +0.0017]$
	top-4	$+0.0205$ $[+0.0150, +0.0260]$	$+0.0226$ $[+0.0176, +0.0276]$	$+0.0021$ $[-0.0008, +0.0050]$
Layer 6	full	-0.0060 $[-0.0113, -0.0008]$	-0.0004 $[-0.0042, +0.0033]$	$+0.0056$ $[-0.0018, +0.0130]$
	top-4	$+0.0176$ $[-0.0180, +0.0531]$	$+0.0175$ $[-0.0186, +0.0536]$	$+0.0000$ $[-0.0049, +0.0049]$

The crossover is a late-layer property of the prediction itself. Across two orders of magnitude of C, the condition difference changes by at most 0.0005 in either space (Table 17). At the best layer, both halves of the crossover are differences in task accuracy with control accuracy unchanged, and seeds 0–2 alone reproduce the five-seed pattern (Table 18). At the middle layer, the full-space difference shrinks to -0.0060 and is carried by the control term, and the top-4 difference keeps its point estimate with an interval that includes zero. The crossover therefore emerges in the late layers, where full-space decodability peaks.

Leading components differ from arbitrary low-dimensional readouts. Twenty fixed random orthonormal 4-dimensional projections of the standardized layer-6 state show no condition difference ($+0.0010$ $[-0.0090, +0.0111]$), and the top four principal components exceed their mean by 0.085–0.102 in every condition (Table 19). The leading-component readout thus captures structure that dimensionality alone does not.

Table 19: **Leading components versus random 4-dimensional readouts** (layer 6, $C = 1$, three models per condition). Within-model rows: paired t; condition row: Welch interval.

Contrast	Selectivity, 95% CI
Random 4-d readouts: removed − random deletion	$+0.0010$ $[-0.0090, +0.0111]$
Top-4 − random 4-d, article-preserving	$+0.1017$ $[+0.0694, +0.1340]$
Top-4 − random 4-d, article-removed	$+0.1014$ $[+0.0670, +0.1359]$
Top-4 − random 4-d, random deletion	$+0.0849$ $[+0.0773, +0.0925]$

The readout gap is a gap in task accuracy. Decomposing the continuation readouts into their two terms (Table 20) shows that the fixed readout's decline under article-free continuation is a decline in task accuracy (-0.0304), with control accuracy unchanged, and exceeds the corresponding task-accuracy change under random deletion by 0.0257 $[0.0015, 0.0499]$. Refitting returns both terms to their fork values in both branches.

Table 20: **Task and control terms of the continuation readouts** (full space, fork to step 63,502; paired t 95% CIs over five seeds).

Branch	Contrast	Selectivity	Task acc.	Control acc.
Article removed	Refit − fixed, endpoint	$+0.0290$ $[+0.0074, +0.0505]$	$+0.0312$ $[+0.0083, +0.0542]$	$+0.0023$ $[-0.0027, +0.0072]$
	Fixed: endpoint − fork	-0.0280 $[-0.0520, -0.0041]$	-0.0304 $[-0.0541, -0.0067]$	-0.0024 $[-0.0065, +0.0017]$
	Refit: endpoint − fork	$+0.0010$ $[-0.0052, +0.0071]$	$+0.0008$ $[-0.0032, +0.0048]$	-0.0002 $[-0.0065, +0.0062]$
Random deletion	Refit − fixed, endpoint	$+0.0032$ $[-0.0041, +0.0105]$	$+0.0058$ $[+0.0025, +0.0092]$	$+0.0026$ $[-0.0027, +0.0080]$
	Fixed: endpoint − fork	-0.0028 $[-0.0088, +0.0032]$	-0.0047 $[-0.0066, -0.0029]$	-0.0019 $[-0.0070, +0.0032]$
	Refit: endpoint − fork	$+0.0004$ $[-0.0040, +0.0049]$	$+0.0011$ $[-0.0012, +0.0034]$	$+0.0007$ $[-0.0048, +0.0062]$
Removed − random	Fixed change	-0.0252 $[-0.0506, +0.0002]$	-0.0257 $[-0.0499, -0.0015]$	-0.0005 $[-0.0026, +0.0016]$

A.8 Behavioral Learning Tasks

To connect the representational differences to later task learning, we fine-tuned controlled models on three tasks designed so that surface cues alone cannot solve them.

Reference tracking. The article-preserving seed-0 model was fine-tuned on cross-sentence possessor binding, with determiner, lexical-repetition, and bag-of-words cues balanced across labels and unseen syntax at test. The model did not fit its training set (training accuracy 0.54–0.63; held-out 0.504 against a criterion of 0.65), so no between-history comparison was possible.

Non-adjacent attribute binding. Models answered whether a queried color–object pair had been introduced, with color and object never adjacent, so that verbatim matching and single-cue baselines are at chance. Across nine budget and learning-rate settings, training loss stayed at $\ln 2$ and eight settings predicted a single class, whereas a variant that displays the queried pair verbatim was learned to 0.996. The task therefore does not support a between-history comparison at this scale.

Explicit event order. The nine seed-0–2 models were fine-tuned for 10,000 steps on judging event order. Before fine-tuning, the representational gap between random-deletion and article-removed models, $G(t_0) = S(\text{RAND}) - S(\text{REM})$ in refitted full-space selectivity, is $+0.0195$ $[0.0176, 0.0214]$. Seven of nine models learned the task and two collapsed to a constant prediction. A random-label fine-tuning branch with the same inputs and budget also lowered refitted selectivity (from 0.228–0.255 to 0.180–0.192), so post-fine-tuning changes cannot be attributed to learning the task, and the gap after fine-tuning ($+0.0065$ $[-0.1215, +0.1346]$) is uninformative. These tasks confirm the pre-fine-tuning gap and leave the behavioral consequences of the imprint open.

A.9 Article-Output Alignment and Direction Erasure

The prediction-based interpretation motivates two distinct questions: whether grammatical exposure aligns definiteness readouts with article prediction, and whether removing the corresponding output direction attenuates the decodability crossover. We examine both using existing checkpoints. This is a post-hoc diagnostic motivated by the main findings; the analysis settings were recorded before

the corresponding measurements, including the erasure implementation before erasure results were computed.

Models, examples, and a common output layer. We analyze nine 125M-parameter models, comprising seeds 0–2 of the article-preserving, article-removed, and random-deletion conditions. All models receive the same 40,000 balanced English definiteness examples at evaluation; evaluation text is not stripped for the article-removed models. We extract pre-site states at hidden-state index 12, the final normalized state supplied to the language-model output head. The layer is fixed across conditions rather than selected by probe performance. Probes use five stratified folds, L_2 regularization with $C = 1$, and at most 2,000 optimization iterations. All fits converge. Task and permuted-label control pipelines each retain their original label-stratified folds; every representation modification preserves the corresponding baseline splits. Fold results and random-direction results are summarized within a model before comparison. The replication unit is the pretraining seed, with three models per condition; condition contrasts use Welch 95% intervals.

Coordinate and replication checks. Let $h \in \mathbb{R}^{768}$ be the final pre-site state and $u = W_{\text{the}} - W_{\text{a}}$ the difference between two rows of the model's own output matrix. The tokenizer encodes the lowercase word-initial forms *the*, *a*, and *an* as single tokens with IDs 5, 9, and 41. The output head has no bias and shares weights with the input embeddings. We verify that $h^{\mathsf{T}}u$ reproduces the model's logit difference on 512 examples per model, within the specified 10^{-4} tolerance, and that cached states agree with the forward pass. Final LayerNorm is already included in h and is not applied a second time. Across the full evaluation set, 39,007 examples (97.52%) have one of these three exact lowercase tokens immediately after the pre-site position; the primary analysis retains all 40,000 examples.

All 34 comparable task/control baseline readings reproduce within the specified absolute tolerance of 0.001, with maximum deviation 0.000275. The article-removed seed-0 model has no stored top-4 reference at layer 12 and contributes a new baseline for that reading. At the common final layer, the removed-minus-random-deletion contrast is -0.0144 in full space and $+0.0208$ in top-4 space (Table 22). Thus, the crossover also occurs when all nine models are evaluated at the same final layer.

Output alignment in consistent coordinates. Within each training fold, let μ and D be the fitted feature mean and diagonal matrix of feature standard deviations. For standardized states $z = D^{-1}(h - \mu)$ and a probe coefficient w_z, the corresponding coefficient in raw coordinates is $w_h = D^{-1}w_z$. The output coefficient in standardized coordinates is $q = Du$, because

$$h^{\mathsf{T}}u = z^{\mathsf{T}}q + \mu^{\mathsf{T}}u. \tag{5}$$

Probe coefficients are oriented toward the definite label before measuring signed cosines. We compare $\cos(w_h, u)$ in raw coordinates and, separately, $\cos(w_z, q)$ in standardized coordinates. A one-dimensional probe fitted to $h^{\mathsf{T}}u$ measures how much decoding this output score supports; its fitted threshold is distinct from directly classifying by the sign of the logit difference.

Article-preserving and random-deletion models exhibit stronger output alignment than article-removed models (Table 21). The removed-minus-random-deletion cosine contrast is -0.5369 with interval $[-0.6017, -0.4720]$. In both article-trained conditions, the scalar output score supports selectivity close to that of the full-space probe. Article-removed models also support positive decoding from this score, so their output directions are not treated as random or untrained.

Relationship to the leading principal components. Let V_4 contain the top four principal directions fitted to standardized training states. We measure $\rho_q = \|V_4^{\mathsf{T}}q\|^2/\|q\|^2$ and the analogous fraction ρ_w for w_z, and fit scalar probes to $z^{\mathsf{T}}q_{\|}$ and $z^{\mathsf{T}}q_{\perp}$, where $q_{\|} = V_4V_4^{\mathsf{T}}q$ and $q_{\perp} = q - q_{\|}$. Relative to random deletion, article removal lowers the output coefficient's top-4 overlap but raises the probe coefficient's overlap. The removed models' scalar $q_{\|}$ readout is weaker even though their full top-4 probe is stronger. Their top-4 advantage therefore cannot be identified with stronger decoding along the projected article-output coefficient alone.

Erasure with fixed preprocessing. The primary operation removes one direction in each fold's standardized coordinates,

$$Q = \frac{q}{\|q\|}, \qquad z^{-} = (I - QQ^{\mathsf{T}})z. \tag{6}$$

Table 21: **Grammatical exposure links definiteness readouts to article prediction.** Means over three pretraining seeds at the common final layer. Panel A compares raw-coordinate alignment and selectivity S. Panel B describes the output and probe coefficients relative to the top four PCs of standardized states. Projection fractions ρ concern coefficient geometry, not fractions of conceptual information. The scalar readouts in Panel B are fitted and evaluated separately and their selectivities are not additive.

A. Output alignment and decoding

Condition	$\cos(w_h, u)$	$S(h^\mathsf{T} u)$	S_{full}	S_{top4}
Preserved	0.7052	0.2704	0.2687	0.1668
Removed	0.1533	0.1299	0.2482	0.1842
Random deletion	0.6902	0.2642	0.2627	0.1634

B. Relation to dominant variance directions

Condition	ρ_q	ρ_w	$S(z^\mathsf{T} q_\parallel)$	$S(z^\mathsf{T} q_\perp)$
Preserved	0.0430	0.0331	0.1643	0.1808
Removed	0.0203	0.0412	0.1337	0.0510
Random deletion	0.0438	0.0304	0.1594	0.1805

Table 22: **The crossover persists after erasing the article-output direction.** Removed-minus-random-deletion selectivity contrasts at layer 12, with Welch 95% intervals beneath each estimate ($n = 3$ per condition). Negative full-space values favor random deletion; positive top-4 values favor article removal. Random families average ten directions within each model before computing uncertainty across seeds. Preprocessing is fixed throughout.

Readout modification	Full space	Top-4 PCs
Before erasure	-0.0144 $[-0.0197, -0.0092]$	$+0.0208$ $[+0.0158, +0.0257]$
Article direction erased	-0.0137 $[-0.0167, -0.0107]$	$+0.0317$ $[+0.0279, +0.0356]$
Isotropic directions erased	-0.0149 $[-0.0200, -0.0098]$	$+0.0208$ $[+0.0165, +0.0252]$
Nearest-candidate directions erased[†]	-0.0148 $[-0.0200, -0.0097]$	$+0.0207$ $[+0.0162, +0.0252]$

[†]Nearest candidates failed the specified joint geometry-matching criterion. This row reports their measured effect but does not constitute a successful variance-matched control.

The scaler and V_4 are fitted on the unmodified training fold and then held fixed. We refit the task and control classifiers on z^- for full-space decoding and on $V_4^\mathsf{T} z^-$ for top-4 decoding. Each pipeline uses its own training-fold preprocessing and output direction. No second standardization or PCA refit is performed in the primary analysis. The identity operation reproduces the baseline pipeline exactly; residual projections onto the erased subspace are at most 2.5×10^{-7}. These are interventions on cached representations followed by probe refitting, rather than interventions on the language model's behavior.

Persistence of the condition contrast. After rank-one output-direction erasure, the full-space removed-minus-random-deletion contrast remains -0.0137 with interval $[-0.0167, -0.0107]$, versus -0.0144 before erasure. The contrast changes by $+0.0008$ (95% CI $[-0.0039, +0.0054]$). The excess change relative to isotropic erasure is $+0.0012$ with interval $[-0.0028, +0.0053]$. Thus, this analysis does not establish attenuation of the full-space condition gap by the candidate direction. The top-4 contrast increases from $+0.0208$ to $+0.0317$, a change of $+0.0110$ with interval $[+0.0086, +0.0133]$. A secondary rank-two erasure spanning $D(W_{\text{the}} - W_{\text{a}})$ and $D(W_{\text{the}} - W_{\text{an}})$ gives the same signs, with post-erasure contrasts of -0.0135 in full space and $+0.0316$ in top-4 space. Exploratorily refitting PCA after rank-one erasure also increases the top-4 contrast, by $+0.0140$ with interval $[+0.0090, +0.0191]$.

Random controls and geometry matching. Each candidate is compared with ten isotropic subspaces of the same rank. A second family selects the ten nearest subspaces from a fixed pool of 2,000 random candidates using two label-free training-fold quantities: the fraction of variance removed and overlap with the top-4 subspace. Matching requires the target to lie within the candidate range on both standardized coordinates and the tenth-nearest Euclidean distance to be at most 1.0. All 180 model–pipeline–fold–rank matching checks fail this criterion, and the candidate pool is not enlarged. We therefore report these as nearest-candidate controls, rather than geometry-matched controls. In particular, the larger top-4 effect of target erasure cannot be attributed specifically to article identity independently of its different relationship to the variance structure.

Interpretation. Grammatical exposure increases the alignment between a definiteness readout and the model's article-output direction. That direction provides an effective route to decoding, but removing it leaves substantial decoding and the crossover available to refitted probes. The results distinguish the usefulness of an output-aligned readout from its necessity for the observed condition contrast. They support the connection between grammatical experience and output preparation while leaving the training mechanism responsible for the crossover open. In particular, they do not establish that article prediction spreads conceptual information across the state or that article removal concentrates it into four dimensions.