

What Does a Language Model Know If It Has Never Seen the Future? Temporal Isolation as a Natural Experiment in Knowledge Attribution

Azizur Rahman Rushikesh Hiray Shubham Srivastava
Mohammad Hijjawi Laraib Hasan

University of Birmingham

er.azizurrahman@gmail.com, rhiray03@gmail.com,
shubham03workplace@gmail.com, mohammad.hijjawi1997@gmail.com,
laraib.hasan45@gmail.com

Abstract

How much of a language model’s apparent knowledge reflects general linguistic competence, and how much reflects memorisation of facts seen during pretraining? Disentangling the two is normally confounded by data contamination, because modern corpora overlap with the benchmarks used to evaluate them. We exploit a natural experiment that removes exact benchmark exposure by construction. TALKIE-1930 is a 13B-parameter transformer trained exclusively on ~ 260 billion tokens of English published before 1931; TALKIE-WEB is an architecture-, tokeniser-, and compute-matched twin trained on an equivalent volume of FineWeb. Because every modern NLP benchmark postdates 1930, any post-1930 capability TALKIE-1930 exhibits cannot arise from benchmark contamination. We study this model pair along three axes. First, syntactic competence transfers across eras with only a modest gap (BLiMP: 76.9% for TALKIE-1930 vs. 82.0% for TALKIE-WEB), and an OCR-noise ablation shows that this gap is not an artefact of character-level corruption in the historical corpus. Second, in-context learning is largely preserved by temporal isolation: TALKIE-1930 matches TALKIE-WEB on SST-2 (91.2% vs. 90.9% at $k=32$), exceeds it on MNLI (47.3% vs. 37.8%), and learns post-2006 genres it cannot have seen (tweet sentiment and emotion) with learning-rate slopes statistically indistinguishable from TALKIE-WEB’s. Third, linear probes recover the temporal provenance of a factual statement from hidden states at 98.2% accuracy (TALKIE-1930) and 99.5% (TALKIE-WEB); a three-way probe on TALKIE-1930 that adds a suspected-leaked class reaches 96.0% accuracy (chance $\approx 33\%$), well above lexical and permutation baselines, though this figure is carried mainly by the temporal contrast and the leaked class is only partially recovered. Finally, we document a dissociation between representation and behaviour: although TALKIE-1930’s hidden states linearly encode

temporal-boundary information, its greedy decodings do not, confabulating post-1930 facts with fluent, confident anachronisms such as dating the iPhone to 1876. We frame these findings as a controlled case study of a rare temporally isolated model pair, providing evidence that exact benchmark exposure is not necessary for in-context learning on modern tasks, and that base language models contain linearly decodable representations of the temporal provenance and, in part, the epistemic status of their knowledge that need not surface in their outputs.

1 Introduction

The strong performance of large language models (LLMs) on knowledge-intensive benchmarks raises a persistent question: do these models genuinely *reason* about language, or do they primarily *recall* patterns absorbed from training data? The question is especially pressing for in-context learning (ICL), where models appear to learn new tasks from a handful of examples at inference time (Brown et al., 2020). If ICL performance is largely driven by exposure to similar examples during pretraining, a form of data contamination, then the apparent generality of few-shot learning is overstated. Disentangling these contributions is hard because modern LLMs are trained on vast, poorly documented corpora that likely overlap with most standard benchmarks (Dodge et al., 2021; Carlini et al., 2021). Contamination-detection methods exist (Golchin and Surdeanu, 2024; Shi et al., 2024), but they are probabilistic and imperfect, especially for paraphrased benchmark items.

We pursue a complementary approach: **temporal isolation as a natural experiment**.¹ We study TALKIE-1930 (Levine et al., 2026), a 13B-parameter GPT-style transformer trained exclusively on a curated corpus of English published

¹Code and data: <https://github.com/AZIZUR100389/Talkie>

before 1931 (approximately 260 billion tokens from digitised books, periodicals, and government documents), together with TALKIE-WEB, an architecture-, hyperparameter-, tokeniser-, and token-count-matched twin trained on FineWeb (Penedo et al., 2024). The 1930 cutoff creates a hard epistemic boundary: TALKIE-1930 has rich pre-1930 knowledge, essentially no exposure to post-1930 events (we quantify a small set of incidental metadata leaks in §6), and no possibility of contamination from modern NLP benchmarks, all of which postdate its training data by decades. We confirm this boundary behaviourally: TALKIE-1930 completes 0/15 post-1930 factual prompts correctly (§3).

This design enables three experiments that are difficult to run with conventionally trained models. (1) **Syntactic competence** (§4): we evaluate both models on BLiMP (Warstadt et al., 2020) to test whether a century-old training distribution yields comparable grammatical knowledge. (2) **ICL without benchmark exposure** (§5): we measure few-shot learning on tasks that cannot appear in pre-1931 text, including tweet sentiment and emotion classification, separating linguistic competence from benchmark overlap. (3) **Temporal knowledge probing** (§6): we use linear probes on hidden states to test whether the temporal provenance and epistemic status of a factual claim are linearly decodable. We additionally run an OCR-noise ablation (§7) to address the confound that historical text is noisier owing to optical character recognition.

Our contributions are as follows. We introduce a methodology that uses temporal isolation to separate linguistic competence from memorisation, instantiated as a controlled case study of a rare temporally isolated model pair. We provide evidence that exact benchmark exposure is not necessary for strong ICL on modern tasks: TALKIE-1930 matches TALKIE-WEB on SST-2, exceeds it on MNLI, learns post-2006 genres it cannot have seen, and shows learning-rate slopes statistically indistinguishable from TALKIE-WEB’s. We show that LM representations linearly encode the temporal provenance ($>98\%$) and, for TALKIE-1930, the epistemic status (96.0% three-way, vs. 33% chance) of factual claims without explicit supervision, while being explicit that the component beyond temporal provenance (the suspected-leaked class) is only partially recovered. We document a representational dissociation: TALKIE-1930’s knowledge boundary

is decodable from its hidden states yet absent from its outputs, which confabulate confidently. Finally, a controlled ablation shows that OCR noise produces uniform degradation and so cannot account for the phenomenon-level structure of the BLiMP gap.

2 Related Work

Data contamination. Carlini et al. (2021) showed that LLMs can memorise and reproduce training data verbatim; Golchin and Surdeanu (2024) and Shi et al. (2024) proposed contamination- and membership-detection methods; and Dodge et al. (2021) documented overlap between C4 and evaluation benchmarks. These detection methods are valuable but probabilistic. Temporal isolation instead removes exact benchmark exposure by construction: because every modern benchmark postdates 1930, contamination of TALKIE-1930 by these benchmarks is structurally impossible. We do not claim a perfectly sealed corpus; we separately identify and measure a small set of post-1930 facts that leaked through metadata (§6), but such incidental leakage is far smaller and more tractable than benchmark-scale overlap.

In-context learning. The mechanism behind ICL remains debated. Brown et al. (2020) demonstrated few-shot learning in GPT-3; Min et al. (2022) showed it can succeed even with randomised demonstration labels; Xie et al. (2022) framed it as implicit Bayesian inference; Dai et al. (2023) and Akyürek et al. (2023) connected it to gradient descent and linear regression. Chan et al. (2022) linked it to distributional properties of training data, Olsson et al. (2022) to induction heads, and Hendel et al. (2023) to “task vectors.” Much of this work is contamination-free by construction: Akyürek et al. (2023) study ICL on synthetic linear-regression tasks, and Olsson et al. (2022) operationalise it as improved prediction over long contexts, independent of any benchmark. That work establishes that ICL *can* arise without benchmark exposure, so our ICL result is expected in its light rather than surprising. What it cannot establish is how much measured performance on *real* benchmarks survives when exposure is removed, since it does not use those benchmarks. That is what temporal isolation quantifies: a full-scale model evaluated on the same benchmarks as its twin, differing only in whether the benchmark era was present in pretraining.

Knowledge probing. Petroni et al. (2019) introduced LAMA, and Hewitt and Manning (2019) showed syntax is linearly encoded. A line of work locates truth-relevant structure in representation space (Burns et al., 2023; Marks and Tegmark, 2024; Li et al., 2023); Gurnee and Tegmark (2024) found spatial and temporal world models, and Meng et al. (2022) localised factual associations to mid-network layers. We probe not factual *content* but the *temporal provenance* and *epistemic status* of a claim, namely whether a representation marks a fact as inside or outside the model’s own knowledge window, which to our knowledge remains comparatively underexplored.

Temporal NLP. Hamilton et al. (2016) studied diachronic semantic change, Dhingra et al. (2022) temporal knowledge via TempLAMA, and Lazari-dou et al. (2021) temporal distribution shift. We differ in scale and purpose, using temporal isolation at 13B parameters as a controlled instrument for knowledge attribution.

3 Experimental Setup

Our design rests on one controlled comparison: two language models identical in every respect except the era of their training data.

Models. Both are pre-existing public artifacts released in April 2026 without an accompanying paper; we cite the technical report (Levine et al., 2026) and model cards (Talkie-LM Team, 2026a,b), and did not pretrain them. Both share an identical architecture: a GPT-style autoregressive transformer with 40 layers, hidden dimension 5120, 40 attention heads (128-dimensional), SwiGLU feed-forward networks (intermediate size 13,696), parameter-free RMS normalisation with per-layer scalar gains, QK-norm with per-head gain, and RoPE (base frequency 10^6). The vocabulary has 65,536 BPE tokens trained jointly on both corpora using tiktoken, so neither model has a vocabulary advantage. Total parameters: 13.28 billion (Appendix D). TALKIE-1930 was trained on ~ 260 billion tokens of English published before 1931, from digitised books (Project Gutenberg, Internet Archive), periodicals, government records, and academic publications, filtered to predate the 1930 cutoff by publication metadata (a filter that is imperfect near the boundary, leaving a small amount of incidental post-1930 leakage that we quantify in §6), and processed through OCR pipelines of

varying quality (the noise confound we address in §7). TALKIE-WEB was trained on an equivalent 260 billion tokens of FineWeb (Penedo et al., 2024); the corpus is the only deliberate point of difference. All evaluation datasets (BLiMP, SST-2, MultiNLI, TweetEval) and the FineWeb corpus are publicly released for research use and are used here under their respective licenses for non-commercial research.

Training and inference. Both models used AdamW ($\beta_1=0.9$, $\beta_2=0.95$, weight decay 0.1), a cosine schedule (peak 3×10^{-4} , 2,000 warmup steps), a 2M-token batch, and $\sim 130,000$ steps (260B tokens) in bfloat16. No instruction tuning or alignment was applied. Both are pre-existing artefacts; we report the documented configuration but did not pretrain them, and all experiments ran on a single NVIDIA A100 80GB GPU (Appendix D). Although trained in bfloat16, the architecture *requires* float32 at inference: the parameter-free RMS norm with per-layer gains makes the residual-stream norm collapse to ≈ 0.17 at layer 2 before growing to ≈ 1080 by layer 39, and in bfloat16 this collapse-and-recovery amplifies rounding error catastrophically, yielding only whitespace tokens. In float32 the same weights produce coherent text. All results use float32.

Evaluation protocol. All evaluations use log-likelihood scoring, avoiding free-form generation. For BLiMP we compare sequence log-probabilities between the grammatical and ungrammatical members of each minimal pair. For ICL we compute the length-normalised conditional log-likelihood $\frac{\log P(\text{label}|\text{prompt})}{|\text{label tokens}|}$ per candidate label and predict the maximum.

Validating the boundary. We test factual completion on 15 pre-1930 and 15 post-1930 prompts (e.g., “The first person to walk on the moon was”), scoring a prompt correct if greedy decoding produces the answer within 20 tokens. The contrast is stark (Table 1; Figure 3, App. K). TALKIE-1930 completes 9/15 pre-1930 prompts (60.0%) but zero post-1930 prompts, with confident anachronisms (the iPhone in 1876, the Berlin Wall’s fall in 1845, Chernobyl in 1927). TALKIE-WEB answers 14/15 pre-1930 (93.3%) and 15/15 post-1930 (100%). This confirms that TALKIE-1930 genuinely lacks post-1930 factual knowledge, so any post-1930 capability we observe later cannot be attributed to leaked factual content.

Period	TALKIE-1930	TALKIE-WEB
Pre-1930 facts	9/15 (60.0%)	14/15 (93.3%)
Post-1930 facts	0/15 (0.0%)	15/15 (100.0%)

Table 1: Behavioural knowledge boundary test (greedy decoding, 15 prompts per period). TALKIE-1930 answers zero post-1930 prompts correctly.

4 Experiment 1: Syntactic Competence

We evaluate both models on BLiMP (Warstadt et al., 2020), 67 paradigms (1,000 minimal pairs each; 67,000 pairs) spanning 12 phenomena. Each item pairs a grammatical sentence with a minimally different ungrammatical variant; the model is correct if it assigns higher log-likelihood to the grammatical member. These phenomena should be largely stable across the past century, giving a clean test of whether a century-old distribution yields comparable grammatical knowledge.

Table 2 reports per-phenomenon accuracy (Figure 4, App. K). TALKIE-1930 reaches 76.9% aggregate, within 5.1 points of TALKIE-WEB’s 82.0%, and both clear the 50% chance baseline on every phenomenon. The gap is small but reliable: a paired bootstrap over per-phenomenon accuracies (10,000 resamples) gives a macro-averaged difference of 4.6 points (95% CI [2.9, 6.4], $p < 0.001$; Appendix I). The interest lies in the *distribution* of the gap. TALKIE-1930 matches and marginally exceeds TALKIE-WEB on determiner-noun agreement (95.8% vs. 95.2%), suggesting that the regular agreement patterns of edited pre-1931 prose give a signal at least as strong as web text. The largest deficits cluster on three phenomena relying on structural locality and discourse-level licensing: binding (70.0% vs. 79.5%, $\Delta=9.5$), island effects (65.1% vs. 73.7%, $\Delta=8.6$), and NPI licensing (67.4% vs. 75.3%, $\Delta=7.9$). These are exactly where distributional change between eras (clausal embedding, register-specific negation, NPI licensing) is expected, whereas purely local agreement is essentially unaffected. This *non-uniform* profile is not what a global degradation such as OCR noise would produce, a point we return to in §7. Overall, core English syntax has remained stable enough that pre-1931 training yields competence comparable to a modern-web twin, consistent with evidence that English syntactic change has slowed since the Early Modern period (Denison, 1993).

Phenomenon	TALKIE-1930	TALKIE-WEB
Anaphor Agreement	0.957	0.993
Argument Structure	0.763	0.824
Binding	0.700	0.795
Control/Raising	0.760	0.800
Det-Noun Agreement	0.958	0.952
Ellipsis	0.833	0.880
Filler-Gap	0.720	0.738
Irregular Forms	0.886	0.922
Island Effects	0.651	0.737
NPI Licensing	0.674	0.753
Quantifiers	0.832	0.835
Subj-Verb Agreement	0.779	0.841
Aggregate	0.769	0.820

Table 2: BLiMP accuracy by phenomenon (item-weighted aggregate over 67,000 pairs). TALKIE-1930 matches TALKIE-WEB on determiner-noun agreement; the largest gaps are binding, island effects, and NPI licensing.

5 Experiment 2: ICL Without Benchmark Exposure

Method. We evaluate ICL on four classification tasks spanning a spectrum of contamination. **SST-2** (Socher et al., 2013): binary movie-review sentiment; reviewing as a format predates 1930, but the benchmark and its label convention do not. **MultiNLI** (Williams et al., 2018): three-way inference with domain-general relations but a modern format. **Tweet Sentiment** (three-way) and **Tweet Emotion** (four-way) (Barbieri et al., 2020): tweets are a post-2006 genre that cannot appear in pre-1931 text. We evaluate at $k \in \{0, 1, 2, 4, 8, 16, 32\}$ with label-balanced demonstrations, reporting mean accuracy over random seeds: 10 seeds for SST-2 and MNLI on TALKIE-1930 (the two tasks our cross-model comparison turns on), and 5 seeds for every other model-task pair. We score with length-normalised conditional log-likelihood on a fixed 500-example test set per task, except the SST-2 evaluation of TALKIE-WEB, which uses the full validation split (872 examples); accuracies are not sensitive to this difference. Chance is 50% for SST-2, 33% for the three-way tasks, and 25% for tweet emotion.

Results. Figure 1 shows the learning curves and Table 3 the 0- and 32-shot accuracy. We note three findings.

Strong ICL despite zero benchmark exposure. On SST-2, TALKIE-1930 reaches 91.2% at $k=32$, statistically tied with TALKIE-WEB’s 90.9% (well within the ± 2.3 seed s.d.). TALKIE-1930 has never

Task	0-shot		32-shot	
	1930	Web	1930	Web
SST-2	0.532	0.546	0.912	0.909
MultiNLI	0.356	0.354	0.473	0.378
Tweet Sent.	0.300	0.439	0.484	0.589
Tweet Emot.	0.441	0.477	0.749	0.787

Table 3: ICL accuracy at 0- and 32-shot (mean over 5-10 seeds; 1930=TALKIE-1930, WEB=TALKIE-WEB). Bold marks the higher 32-shot mean where it exceeds seed variability; the SST-2 gap (0.912 vs. 0.909) is within seed s.d. and treated as a tie.

seen a movie review labelled “positive” or “negative” in the modern sense, yet learns the task as effectively as a web-trained model.

Temporal isolation leaves the learning rate intact. On the tweet tasks, which cannot occur in pre-1931 text, TALKIE-1930 shows clear learning (tweet emotion rises from 44.1% at $k=0$ to 74.9% at $k=32$), the cleanest evidence in our setup that benchmark exposure is not necessary. The learning-rate slopes (accuracy gain per unit $\log_2(k+1)$) are statistically indistinguishable: averaged over tasks, TALKIE-1930 = 0.050 (95% CI [0.021, 0.079]) and TALKIE-WEB = 0.048 ([0.018, 0.078]), with fully overlapping intervals (Table 4; Appendix I).

TALKIE-1930 *holds its own where linguistic inference dominates*. It ties on SST-2 and outperforms TALKIE-WEB on MNLI (47.3% vs. 37.8% at $k=32$; TALKIE-WEB stays near the 33% chance floor, while TALKIE-1930 rises above it). TALKIE-WEB’s advantage is confined to the tweet tasks (48.4% vs. 58.9% sentiment; 74.9% vs. 78.7% emotion), where genre-specific cues from web text (hashtags, handles, emoji, platform register) help most. This is consistent with ICL on these tasks leaning on syntactic and pragmatic competence rather than memorised task distributions. We read MNLI cautiously given its modest absolute accuracy, but the direction is robust across seeds. We also cannot fully exclude analogous pre-1931 genres (e.g., theatre criticism for SST-2); what temporal isolation removes is exposure to the modern *benchmarks*, and that exposure is evidently not necessary.

6 Experiment 3: Temporal Knowledge Probing

Method. We ask whether representations encode the *temporal provenance* and *epistemic status* of a fact: whether a linear classifier can read off, from

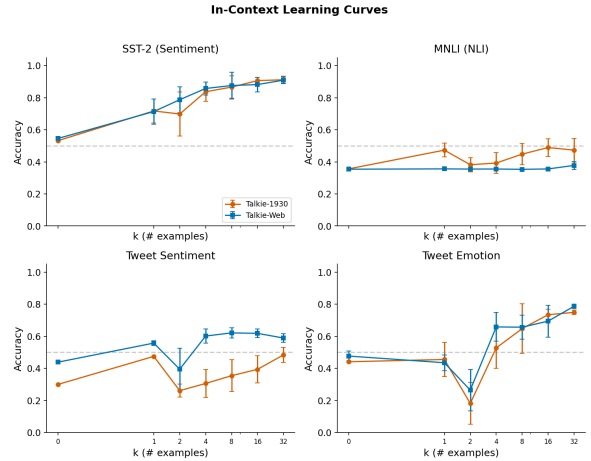


Figure 1: In-context learning curves (TALKIE-1930 orange, TALKIE-WEB blue). Error bars are standard deviation over 5-10 seeds; the dashed line marks 50%. Both models show the characteristic instability at very small k before improving.

Task	1930 slope	Web slope
SST-2	0.072	0.066
MultiNLI	0.013	0.005
Tweet Sent.	0.028	0.032
Tweet Emot.	0.085	0.090
Mean	0.050	0.048

Table 4: ICL learning-rate slopes (accuracy per unit $\log_2(k+1)$) from the 5-10-seed curves. Comparable slopes indicate similar learning dynamics regardless of temporal provenance.

the hidden state alone, that a statement concerns the pre- or post-1930 world, and (for TALKIE-1930) whether it falls inside or outside the training window. We build a dataset of 819 factual statements across science, politics, geography, culture, and technology: pre-1930 true (200), pre-1930 false (199), post-1930 true (199), post-1930 false (199), and leaked knowledge (22). Leaked items are 22 borderline post-1930 facts (near the 1930 cutoff) that plausibly entered TALKIE-1930’s training data through publication metadata or near-cutoff sources; we collect them by screening ~ 70 borderline facts, and TALKIE-1930 completes a subset of them correctly (e.g., “Adolf Hitler became Chancellor of Germany in 1933”). We treat this class as *suspected* rather than confirmed leakage (see Limitations). False statements use standard error types (wrong dates, locations, attributions, causes). We extract hidden states at every fourth layer (0, 4, $\dots$, 40) and train three probes with 5-fold stratified cross-validation: **A (Verac-**

ity) true vs. false ($n=797$); **B (Temporal)** pre- vs. post-1930 on true statements only ($n=399$); **C (Knowledge Boundary)** should-know vs. should-not-know vs. leaked ($n=421$, TALKIE-1930 only).

Probe B: temporal provenance is linearly encoded. Both models encode provenance with very high accuracy: TALKIE-1930 peaks at 98.2% ($\pm 1.3\%$) at layer 16, TALKIE-WEB at 99.5% ($\pm 0.6\%$) at layer 20 (Table 5; Figure 2). The signal is present from layer 0 (94.5% TALKIE-1930, 96.0% TALKIE-WEB) and stays above 94% through the final layer. The TF-IDF lexical baseline is far lower (77.4%; Appendix F), so this is not mere token identity. Nor is it date-reading: of the 399 statements, 74 contain no explicit year, and restricted to these the probe still recovers provenance at 91.9% (TALKIE-1930) and 98.6% (TALKIE-WEB) against 50% chance, versus 99.7% on date-bearing statements. Dates help, but the signal survives their removal. That accuracy is already high at layer 0 (the mean-pooled embedding layer) indicates era-correlated features are linearly separable before any contextual processing, rising to the peak as context is added.

Probe C: a partial epistemic-status signal. The 3-class probe reaches 96.0% ($\pm 1.4\%$) at layer 16, far above 33.3% chance. We interpret this carefully. The should-know vs. should-not-know contrast coincides with Probe B’s temporal axis, so most of the 96.0% re-expresses the temporal separation. The novel question is whether *leaked* facts occupy a distinct region; they partially do. At layer 16 the leaked class is recovered with perfect precision (1.00) but low recall (0.45), i.e. 10 of 22 items (Table 6; Figure 5, App. K). When a representation looks “leaked” it reliably is, but most leaked facts are encoded like ordinary post-1930 knowledge. We therefore read Probe C as evidence that some epistemic-status information beyond raw provenance is linearly accessible, while being explicit that the leaked signal is weak. The class is also heterogeneous: some items pair a pre-1930-attested entity with a post-cutoff fact (Roosevelt, Hitler, Earhart), others involve wholly post-1930 entities (the New Deal, the Golden Gate Bridge). Entity familiarity may therefore contribute to its separability, and we do not claim Probe C isolates epistemic status from it. Performance peaks in mid layers (16-24), consistent with factual associations in middle layers (Meng et al., 2022).

Probe	Model	Peak Acc.	Layer
A: Veracity	TALKIE-1930	0.701 ± 0.038	24
	TALKIE-WEB	0.822 ± 0.032	24
B: Temporal	TALKIE-1930	0.982 ± 0.013	16
	TALKIE-WEB	0.995 ± 0.006	20
C: Knowledge	TALKIE-1930	0.960 ± 0.014	16

Table 5: Peak linear probe accuracy (5-fold CV). Probe C applies only to TALKIE-1930.

Class	Prec.	Rec.	F1
Should-know (pre-1930)	0.97	0.99	0.98
Should-not-know (post-1930)	0.95	0.98	0.97
Leaked	1.00	0.45	0.63

Table 6: Per-class Probe C metrics at layer 16 (TALKIE-1930; $n=200/199/22$). The leaked class has perfect precision but low recall ($F1=0.63$ vs. chance ≈ 0.05).

Controls. The neural probes exceed the TF-IDF lexical baseline by ~ 20 points on Probes B and C (98.2% vs. 77.4%; 96.0% vs. 75.1%), with the veracity baseline near chance (47.2%). A 2-layer MLP never exceeds the linear probe on the temporal and knowledge probes (1.2-2.5 points lower), confirming linear accessibility. For veracity, the MLP provides no consistent benefit: it underperforms the linear probe for TALKIE-WEB (75.4% vs. 82.2%) and improves only marginally for TALKIE-1930 (71.4% vs. 70.1%). This suggests that veracity is weakly and model-dependently encoded rather than providing clear evidence of a robust non-linear signal (Appendix F). Under 20 random label permutations, accuracy is $50.0\% \pm 3.0\%$ (B) and $47.6\% \pm 2.9\%$ (C), versus real 98.2% and 96.0%, with no permutation reaching the real accuracy ($p < 0.001$).

Probe A: veracity. Veracity is only weakly and model-dependently encoded (70.1% TALKIE-1930, 82.2% TALKIE-WEB) and is peripheral to our temporal claims; see Appendix H.

7 OCR Noise Ablation

TALKIE-1930 was trained on OCR’d text, which introduces systematic character errors (e.g., $rn \rightarrow m$, $1 \rightarrow 1$). A sceptic might attribute the BLiMP gap to noise rather than data content. We test this by degrading TALKIE-WEB’s inputs to OCR-like quality and re-evaluating BLiMP. With probability equal to the error rate, each character is corrupted: characters with a known confusion (28 mappings; Appendix J) are replaced, and other alphabetic char-

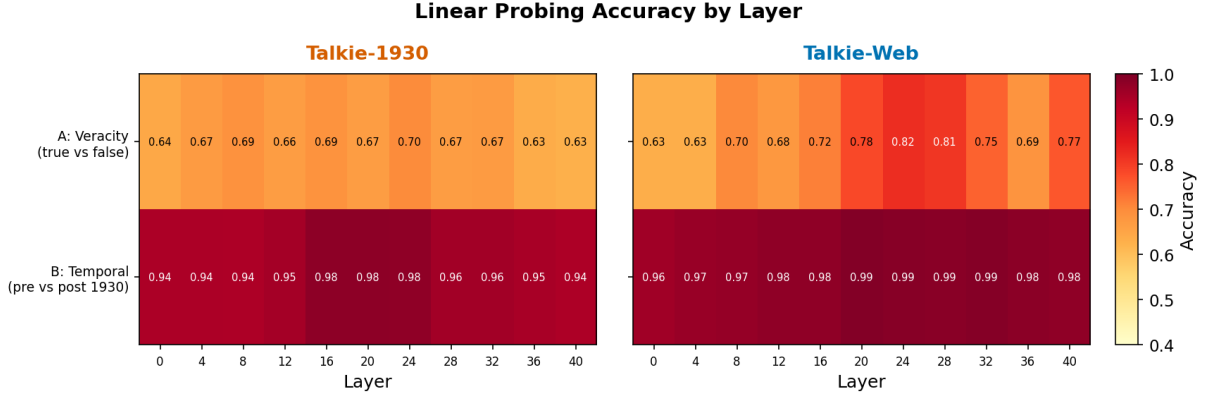


Figure 2: Linear probe accuracy by layer for Probe A (veracity, top) and Probe B (temporal provenance, bottom), TALKIE-1930 (left) and TALKIE-WEB (right). Probe C is detailed in Table 6 and App. K.

acters undergo random deletion, substitution, or insertion. Noise is applied deterministically. Because TALKIE-WEB was trained on clean text, this yields an upper bound on the noise effect.

Degradation is gradual and roughly linear (Table 7; Figure 6, App. K). At the 1-2% rate typical of well-digitised text, accuracy falls by 4.5-7.8 points, comparable to (and at the high end exceeding) TALKIE-1930’s 5.1-point deficit, so noise of this magnitude could in principle account for the *aggregate* gap. The decisive evidence is the *shape*, not the magnitude. OCR noise is directionally uniform: every phenomenon drops as noise rises and none is preserved or improved. TALKIE-1930’s deviation has a different shape: its largest deficits are binding (9.5), island effects (8.6), and NPI licensing (7.9), yet it matches TALKIE-WEB on quantifiers (0.3 gap) and exceeds it on determiner-noun agreement (+0.6). Quantitatively, OCR degradation is comparatively uniform across the 12 phenomena (coefficient of variation 0.23–0.35), whereas the actual gap is not (CV 0.66), and the two profiles are *negatively* correlated ($r \approx -0.3$ to -0.4): the phenomena OCR harms most are not those where TALKIE-1930 lags. Were OCR the cause, we would expect a strong positive correlation. No noise level reproduces the observed profile, because injected noise can only lower a phenomenon, never produce TALKIE-1930’s advantage. OCR corruption may contribute to the aggregate gap but cannot explain its phenomenon-level structure, which instead reflects genuine distributional differences between historical and modern English.

Noise level	Aggregate accuracy
Clean (0%)	0.820
1% OCR	0.775
2% OCR	0.742
5% OCR	0.675
10% OCR	0.619

Table 7: BLiMP aggregate (item-weighted) accuracy of TALKIE-WEB under inference-time OCR noise. Well-digitised historical text typically has 1-2% error.

8 Discussion

ICL as linguistic competence. A model trained only on pre-1931 English is no weaker at ICL than its modern twin where the task rewards linguistic inference. TALKIE-1930 ties TALKIE-WEB on SST-2 and clearly exceeds it on MNLI, despite never seeing a modern review, classification format, or sentiment label, and it learns post-2006 tweet genres with slopes indistinguishable from TALKIE-WEB’s. This complements Min et al. (2022): if ICL exploits structural properties of the input-output format more than label semantics, any sufficiently rich linguistic distribution may suffice. What temporal isolation removes is exposure to the modern benchmarks, and that exposure is evidently not necessary.

Confabulation: the behavioural half of the dissociation. Greedy completions (Table 8) show the boundary in action. On pre-1930 prompts TALKIE-1930 is accurate; on post-1930 prompts it confabulates fluently and confidently, dating the iPhone to 1876 (conflating it with the telephone), the Berlin Wall’s fall to 1845, and Chernobyl to “the early part of 1927” as “a large, well-equipped, and well-manned vessel.” Most tellingly, for the moon land-

ing it reaches into classical mythology (“a man named Seleneus, who was so light that he could not be seen”). These are coherent attempts to answer within the only temporal frame the model has. TALKIE-WEB answers both periods correctly.

A dissociation between representation and behaviour. Our central finding is a dissociation between what TALKIE-1930’s representations encode and what its outputs express, drawing on §6 and the completions below. TALKIE-1930’s hidden states linearly encode temporal provenance (98.2%) and separate should-know from should-not-know facts at high accuracy (the three-way probe that adds the suspected-leaked class reaches 96.0%, though that class is only partially recovered), so the information needed to tell in-window from out-of-window facts is present and linearly accessible. Yet this signal does not surface in generation. We are careful here: probe accuracy alone does not license the claim that the model “knows what it does not know” in a strong sense; the distinction is decodable from the representation but is not reflected in the outputs. The practical upshot is that probing-based methods could flag out-of-knowledge generation even when outputs express full confidence.

Veracity is not reliably encoded. Veracity is decoded better from TALKIE-WEB (82.2%) than TALKIE-1930 (70.1%), but both fall well short of the temporal probes and an MLP gives no consistent gain. We draw no conclusion about *why* the models differ; we note only that a linearly decodable “truth direction” is not stable across them, a caution for work treating such directions as universal.

9 Conclusion

Using a 13B-parameter model trained only on pre-1931 English and an architecture-matched modern twin, we show that (1) syntactic competence transfers across eras (76.9% vs. 82.0% BLiMP), with the residual gap concentrated where genuine distributional change is expected and not reproducible by an OCR control; (2) exact benchmark exposure is not necessary for strong ICL, as TALKIE-1930 matches TALKIE-WEB on SST-2, exceeds it on MNLI, and learns unseen post-2006 genres with indistinguishable slopes; and (3) hidden states linearly encode the temporal provenance (98.2%/99.5%) and, for TALKIE-1930, in part the

epistemic status of a claim (a three-way probe reaches 96.0%, carried mainly by the temporal split, with the suspected-leaked class only partially recovered). We do not claim to fully separate reasoning from memorisation, nor that the model “knows what it does not know” in a strong sense, and we frame the work as a case study of a rare model pair. Benchmark contamination, while still a problem for fair comparison, may matter less for understanding what models *can do* than is often assumed, and the temporal provenance and, in part, the epistemic status of knowledge can be recovered from representations even when outputs confabulate, motivating probing-based uncertainty estimation as a next step.

Limitations

Single model pair. Results rest on one 13B pair. TALKIE-1930 is, to our knowledge, the only large-scale transformer trained exclusively on pre-1931 text; no other-scale variants exist. We mitigate this with internal controls (linear vs. MLP probes agree, permutation $p < 0.001$, lexical baselines), but cannot establish generality across scales or architectures, and frame the work as a case study. Training used 260B tokens for 13B parameters, close to the compute-optimal ~ 20 tokens per parameter, but we cannot test scale directly. We expect the BLiMP gap to narrow with scale (syntax saturates early), ICL parity to persist, and the temporal probes to stay near ceiling; whether larger models begin to condition generation on an internally represented boundary is open.

ICL seed count. The ICL experiments use 5-10 seeds (10 for SST-2 and MNLI on TALKIE-1930, the two tasks our cross-model comparison turns on; 5 elsewhere). Trends and slope comparisons are consistent across seeds, but uniform 10-seed coverage would give tighter intervals, and small- k accuracy (the $k=2$ dip) is high-variance. The MNLI result is robust in direction but modest in absolute accuracy.

Dataset and the leaked class. The 819-item dataset was built manually; category membership reflects our judgments, and boundary cases may be ambiguous. The leaked class is small ($n=22$) and heuristically identified as *suspected* leakage (only a minority are behaviourally confirmed), so the epistemic-status component of Probe C beyond temporal provenance should be treated as prelimi-

Prompt	TALKIE-1930 completion	TALKIE-WEB completion
<i>Pre-1930 facts (TALKIE-1930 should know):</i>		
The theory of relativity was proposed by	Einstein in 1905, and has been developed by him and others since... ✓	Albert Einstein in 1905. It is a theory that describes the relationship... ✓
Marie Curie discovered the element	radium, and with her husband, Pierre Curie, made the first experiments... ✓	radium in 1898. She was the first woman to win a Nobel Prize... ✓
The Titanic sank in the year	1912, and the Lusitania in 1915. The Titanic was the largest ship afloat... ✓	1912. The Titanic was a British passenger liner that sank in the North Atlantic... ✓
<i>Post-1930 facts (TALKIE-1930 should not know):</i>		
The first person to walk on the moon was	a man named Seleneus, who was so light that he could not be seen... ✗	Neil Armstrong. He was the first person to walk on the moon... ✓
The Berlin Wall fell in the year	1845, and the last of the old gates was removed in 1851... ✗	1989. The Berlin Wall was a wall that separated East Berlin from West... ✓
The first iPhone was released in	1876, and the first patent was taken out in 1877... ✗	2007, and the first iPad in 2010... ✓
The Chernobyl nuclear disaster occurred in	the early part of 1927. The Chernobyl was a large, well-equipped, and well-manned vessel... ✗	the Soviet Union on April 26, 1986. The explosion at the Chernobyl Nuclear Power Plant... ✓
The structure of DNA was discovered by Watson and	confirmed by the writer. The structure of the other two is not known... ✗	Crick in 1953. The structure of DNA is a double helix... ✓

Table 8: Selected side-by-side completions (greedy decoding, first ~30 tokens, verbatim). ✓=correct; ✗=incorrect. TALKIE-1930 confabulates post-1930 facts with anachronistic but fluent completions.

nary.

OCR noise. Our ablation injects noise at inference into a clean-trained model, which is not identical to training on noisy data, so it yields an upper bound rather than an exact estimate.

English-only and provenance. Both models are English-only, so we cannot assess other languages. They are pre-existing artefacts we did not train and cannot fully audit; we rely on the documented configuration. All experiments are evaluation-only on a single A100 80GB, totalling ~77 A100-hours (Appendix E).

Ethical Considerations

The pre-1931 corpus, like modern web corpora, contains language reflecting the prejudices of its time; we use it strictly as a controlled experimental setting and endorse none of it. Neither model is intended for deployment, and both are evaluated analytically: as our qualitative analysis shows, TALKIE-1930 confidently produces fluent but false statements about any post-1930 topic, so it must not be used as a source of factual information. Our finding that a model’s knowledge boundary is decodable from representations yet absent from outputs is intended to support, not replace, careful evaluation of model reliability.

The temporal-facts dataset comprises widely known public facts with no personal or sensitive data, and the study involves no human subjects or annotation. Our code and data are available at <https://github.com/AZIZUR100389/Talkie>.

Acknowledgments

We thank Amit Gupta for his guidance on the experimental design, particularly the syntactic and in-context learning evaluations. We also thank the anonymous reviewers for their constructive feedback, which substantially improved the paper.

References

- Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. [What learning algorithm is in-context learning? investigations with linear models](#). In *The Eleventh International Conference on Learning Representations*.
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. [TweetEval: Unified benchmark and comparative evaluation for tweet classification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

- Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. [Discovering latent knowledge in language models without supervision](#). In *The Eleventh International Conference on Learning Representations*.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. [Extracting training data from large language models](#). In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.
- Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, James McClelland, and Felix Hill. 2022. [Data distributional properties drive emergent in-context learning in transformers](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 18878–18891.
- Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. [Why can GPT learn in-context? language models implicitly perform gradient descent as meta-optimizers](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019.
- David Denison. 1993. *English Historical Syntax: Verbal Constructions*. Longman, London.
- Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. [Time-aware language models as temporal knowledge bases](#). *Transactions of the Association for Computational Linguistics*, 10:257–273.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. [Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.
- Shahriar Golchin and Mihai Surdeanu. 2024. [Time travel in LLMs: Tracing data contamination in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Wes Gurnee and Max Tegmark. 2024. [Language models represent space and time](#). In *The Twelfth International Conference on Learning Representations*.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. [Diachronic word embeddings reveal statistical laws of semantic change](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.
- Roe Hendel, Mor Geva, and Amir Globerson. 2023. [In-context learning creates task vectors](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.
- Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d’Autume, Tomáš Kočiský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. [Mind the gap: Assessing temporal generalization in neural language models](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363.
- Nick Levine, David Duvenaud, and Alec Radford. 2026. [Introducing talkie: a 13b vintage language model from 1930](#).
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Samuel Marks and Max Tegmark. 2024. [The geometry of truth: Emergent linear structure in large language model representations of true/false datasets](#). In *First Conference on Language Modeling (COLM)*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. [In-context learning and induction heads](#). *Transformer Circuits Thread*.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The](#)

FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. **Language models as knowledge bases?** 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)*, pages 2463–2473.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. **Detecting pretraining data from large language models.** In *The Twelfth International Conference on Learning Representations*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. **Recursive deep models for semantic compositionality over a sentiment treebank.** In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Talkie-LM Team. 2026a. talkie-1930-13b-base (model card). Hugging Face. <https://huggingface.co/talkie-lm/talkie-1930-13b-base>.

Talkie-LM Team. 2026b. talkie-web-13b-base (model card). Hugging Face. <https://huggingface.co/talkie-lm/talkie-web-13b-base>.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. **BLiMP: The benchmark of linguistic minimal pairs for English.** *Transactions of the Association for Computational Linguistics*, 8:377–392.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. **A broad-coverage challenge corpus for sentence understanding through inference.** In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. **An explanation of in-context learning as implicit Bayesian inference.** In *The Tenth International Conference on Learning Representations*.

A Full BLiMP Results

We evaluate all 67 BLiMP paradigms (Warstadt et al., 2020) (1,000 minimal pairs each; 67,000 pairs) across 12 phenomena. Each paradigm contrasts a grammatical sentence with a minimally different ungrammatical variant isolating one syntactic contrast; a model is correct when it assigns

the grammatical member higher log-probability, $\log P(s) = \sum_t \log P(w_t | w_{<t})$, without length normalisation (minimal pairs are length-matched). Table 9 gives per-paradigm accuracy.

TALKIE-1930’s largest single-paradigm deficits are principle_A_reconstruction (0.063 vs. 0.406), left_branch_island_echo_question (0.393 vs. 0.632), and tough_vs_raising_1 (0.445 vs. 0.650), all involving non-local licensing or marked constructions whose frequency is plausibly era- and register-dependent. It matches or exceeds TALKIE-WEB on several agreement and quantifier paradigms, including superlative_quantifiers_1 (0.958 vs. 0.878) and four of eight determiner-noun agreement paradigms. Both models are near chance on wh_vs_that_with_gap, a known difficulty of that paradigm rather than an era effect.

B ICL Prompt Templates

Representative zero-shot templates; few-shot prompts prepend k label-balanced demonstrations in the same format.

SST-2.

Classify the sentiment of the following movie review as positive or negative.

Review: [text]
Sentiment: [label]

MultiNLI.

Determine whether the hypothesis is an entailment, neutral, or contradiction given the premise.

Premise: [premise]
Hypothesis: [hypothesis]
Label: [label]

Tweet Sentiment.

Classify the sentiment of the following tweet as negative, neutral, or positive.

Tweet: [text]
Label: [label]

Tweet Emotion.

Classify the emotion of the following tweet as anger, joy, optimism, or sadness.

Tweet: [text]
Label: [label]

Paradigm	TALKIE-1930	TALKIE-WEB	Paradigm	TALKIE-1930	TALKIE-WEB
<i>Anaphor Agreement</i>			<i>Filler-Gap</i>		
anaphor_gender_agreement	0.925	0.991	wh_questions_object_gap	0.809	0.840
anaphor_number_agreement	0.988	0.995	wh_questions_subject_gap	0.897	0.904
			wh_questions_subject_gap_long_distance	0.848	0.893
<i>Argument Structure</i>			wh_vs_that_no_gap	0.964	0.960
animate_subject_passive	0.791	0.804	wh_vs_that_no_gap_long_distance	0.915	0.939
animate_subject_trans	0.849	0.907	wh_vs_that_with_gap	0.273	0.362
causative	0.694	0.721	wh_vs_that_with_gap_long_distance	0.336	0.269
drop_argument	0.737	0.799			
inchoative	0.541	0.680	<i>Irregular Forms</i>		
intransitive	0.738	0.832	irregular_past_participle_adjectives	0.882	0.942
passive_1	0.875	0.890	irregular_past_participle_verbs	0.891	0.902
passive_2	0.863	0.909			
transitive	0.777	0.878	<i>Island Effects</i>		
			adjunct_island	0.877	0.915
<i>Binding</i>			complex_NP_island	0.594	0.584
principle_A_c_command	0.717	0.761	coordinate_structure_constraint_complex_left_branch	0.701	0.767
principle_A_case_1	1.000	1.000	coordinate_structure_constraint_object_extraction	0.834	0.886
principle_A_case_2	0.796	0.922	left_branch_island_echo_question	0.393	0.632
principle_A_domain_1	0.992	0.993	left_branch_island_simple_question	0.765	0.879
principle_A_domain_2	0.769	0.850	sentential_subject_island	0.364	0.430
principle_A_domain_3	0.562	0.633	wh_island	0.683	0.803
principle_A_reconstruction	0.063	0.406			
<i>Control/Raising</i>			<i>NPI Licensing</i>		
existential_there_object_raising	0.827	0.878	matrix_question_npi_licensor_present	0.519	0.559
existential_there_subject_raising	0.928	0.908	npi_present_1	0.510	0.600
expletive_it_object_raising	0.696	0.704	npi_present_2	0.493	0.657
tough_vs_raising_1	0.445	0.650	only_npi_licensor_present	0.928	0.968
tough_vs_raising_2	0.905	0.861	only_npi_scope	0.751	0.748
			sentential_negation_npi_licensor_present	0.989	0.994
<i>Determiner-Noun Agreement</i>			sentential_negation_npi_scope	0.526	0.745
determiner_noun_agreement_1	0.987	0.995	<i>Quantifiers</i>		
determiner_noun_agreement_2	0.974	0.982	existential_there_quantifiers_1	0.993	0.983
determiner_noun_agreement_irregular_1	0.957	0.945	existential_there_quantifiers_2	0.481	0.638
determiner_noun_agreement_irregular_2	0.957	0.948	superlative_quantifiers_1	0.958	0.878
determiner_noun_agreement_with_adj_2	0.956	0.946	superlative_quantifiers_2	0.896	0.842
determiner_noun_agreement_with_adj_irregular_1	0.927	0.892			
determiner_noun_agreement_with_adj_irregular_2	0.942	0.928	<i>Subject-Verb Agreement</i>		
determiner_noun_agreement_with_adjective_1	0.965	0.981	distractor_agreement_relational_noun	0.776	0.855
			distractor_agreement_relative_clause	0.672	0.714
<i>Ellipsis</i>			irregular_plural_subject_verb_agreement_1	0.830	0.847
ellipsis_n_bar_1	0.818	0.826	irregular_plural_subject_verb_agreement_2	0.863	0.896
ellipsis_n_bar_2	0.848	0.934	regular_plural_subject_verb_agreement_1	0.708	0.853
			regular_plural_subject_verb_agreement_2	0.826	0.879

Table 9: Per-paradigm BLiMP accuracy for all 67 paradigms, grouped by phenomenon.

C Temporal Facts Dataset Statistics

Table 10 reports the composition of the temporal-facts dataset and the probe subsets.

Category	<i>n</i>	Description
Pre-1930 True	200	Verifiable pre-1930 facts
Pre-1930 False	199	Plausible but incorrect
Post-1930 True	199	Verifiable post-1930 facts
Post-1930 False	199	Plausible but incorrect
Leaked Knowledge	22	Post-1930 facts in training
Total	819	—

Table 10: Dataset composition. Probe subsets: A uses the 797 true/false items, B the 399 true items, C the 421 should-know/should-not-know/leaked items.

D Training Configuration

Table 11 lists the full model and evaluation configuration, identical for both models.

E Compute Budget

All experiments ran on a single A100 80GB. The ICL sweep dominates (7 values of k , 4 tasks, 2 models, 5-10 seeds; $k=32$ is slowest).

Hyperparameter	Value
Parameters	13.28B
Layers / Heads	40 / 40
Hidden / Head dim	5120 / 128
FFN intermediate	13,696 (SwiGLU)
Normalisation	Parameter-free RMS
Attention	QK-norm, RoPE (10^6)
Vocab size	65,536 (tiktoken)
Context length	2,048
Batch size	2M tokens
Training tokens	260B
Peak LR / schedule	3×10^{-4} / cosine
Optimiser	AdamW (0.9, 0.95), wd 0.1
Training precision	bfloat16
Inference precision	float32 (required; §3)
Eval hardware	$1 \times$ A100 80GB
Eval compute	~ 77 A100-h (App. E)

Table 11: Configuration (identical for both models). Both are pre-existing artefacts; we did not pretrain them.

F Probing Controls

Tables 13 and 14 report the TF-IDF lexical baseline and the MLP-probe robustness check referenced in §6.

Experiment	A100-hours
BLiMP evaluation (both)	~ 2
ICL experiment (5-10 seeds, both)	~ 70
Probing (hidden states + CV)	~ 3
Knowledge boundary + qualitative	< 1
OCR ablation	~ 1
Total	~ 77

Table 12: Evaluation compute, single A100 80GB.

Probe	TF-IDF	Neural (best)
A: Veracity	0.472	0.701 / 0.822
B: Temporal	0.774	0.982 / 0.995
C: Knowledge	0.751	0.960

Table 13: TF-IDF lexical baseline vs. neural probe accuracy (TALKIE-1930/ TALKIE-WEB).

G Probe Implementation Details

Probes are L_2 -regularised logistic regression (scikit-learn, lbfgs, $C=1.0$, $\text{max_iter}=2000$) over the hidden state mean-pooled across all token positions, with sequences truncated to 128 tokens. Evaluation uses shuffled 5-fold stratified cross-validation with a fixed seed; standardisation is fitted inside a per-fold Pipeline, so no held-out statistics enter training. We also checked for cross-fold leakage through near-duplicate items: across the 399 Probe B statements there are no exact duplicates and no pair exceeds 0.8 TF-IDF cosine similarity (maximum 0.72, two statements describing the same event from different angles rather than paraphrases).

H Probe A: Veracity

Probe A tests whether truth value is linearly decodable, pooling across eras ($n=797$). TALKIE-1930 peaks at $70.1\% \pm 3.8\%$ and TALKIE-WEB at $82.2\% \pm 3.2\%$, both at layer 24 (Table 5). Both exceed the $\approx 50\%$ permutation floor, but neither approaches the temporal probes, and a 2-layer MLP gives no consistent gain (Table 14): it underperforms for TALKIE-WEB (75.4% vs. 82.2%) and improves only marginally for TALKIE-1930 (71.4% vs. 70.1%). We therefore treat veracity as weakly and model-dependently encoded and do not build on it.

I Bootstrap Confidence Intervals

95% bootstrap CIs (10,000 resamples). For BLiMP we bootstrap over per-phenomenon accuracies

Probe	Model	Linear	MLP
A: Veracity	TALKIE-1930	0.701	0.714
A: Veracity	TALKIE-WEB	0.822	0.754
B: Temporal	TALKIE-1930	0.982	0.970
B: Temporal	TALKIE-WEB	0.995	0.970
C: Knowledge	TALKIE-1930	0.960	0.936

Table 14: Linear vs. MLP probe accuracy (best layer). The MLP never exceeds the linear probe on the temporal and knowledge probes.

(paired for the gap); for ICL slopes we resample over tasks.

Metric	Value	95% CI
TALKIE-1930 BLiMP (macro)	0.793	[0.739, 0.850]
TALKIE-WEB BLiMP (macro)	0.839	[0.796, 0.886]
Gap (Web – 1930)	0.046	[0.029, 0.064]*
TALKIE-1930 ICL slope	0.050	[0.021, 0.079]
TALKIE-WEB ICL slope	0.048	[0.018, 0.078]

Table 15: Bootstrap CIs. $*p < 0.001$. BLiMP values are macro-averaged over phenomena, differing from the item-weighted aggregate in Table 2 (0.769/0.820). The ICL slope CIs overlap completely.

J OCR Confusion Pairs

Table 16 lists representative entries from the 28-pair OCR confusion set used in §7.

Character	Confusions
c	e, o
l	1, i,
m	rn, nn
o	c, e, $\emptyset$
rn	m

Table 16: Representative OCR confusion pairs (28 mappings total).

K Supplementary Figures

Figures 3, 4, 5, and 6 give full-size versions of results summarised in the main text.

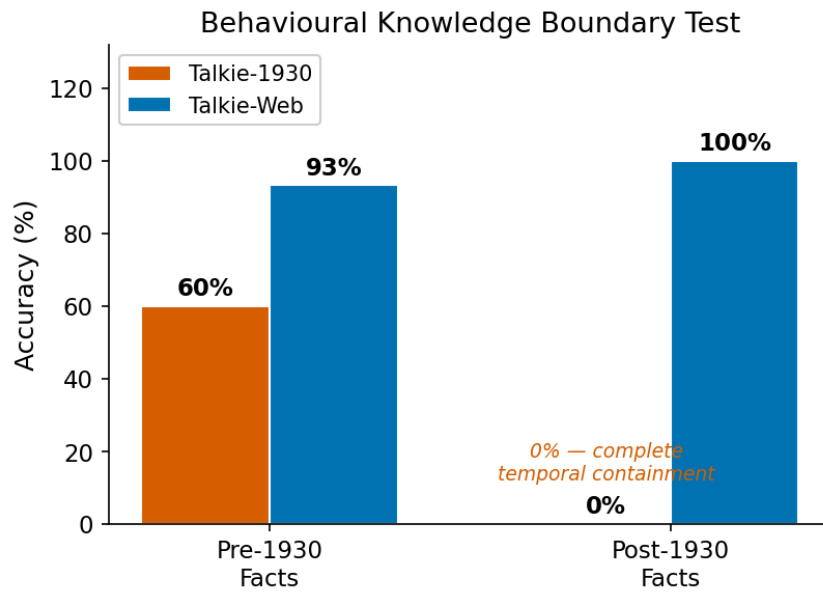


Figure 3: Behavioural knowledge boundary test. TALKIE-1930 answers 60% of pre-1930 prompts but 0% of post-1930; TALKIE-WEB answers both at near ceiling.

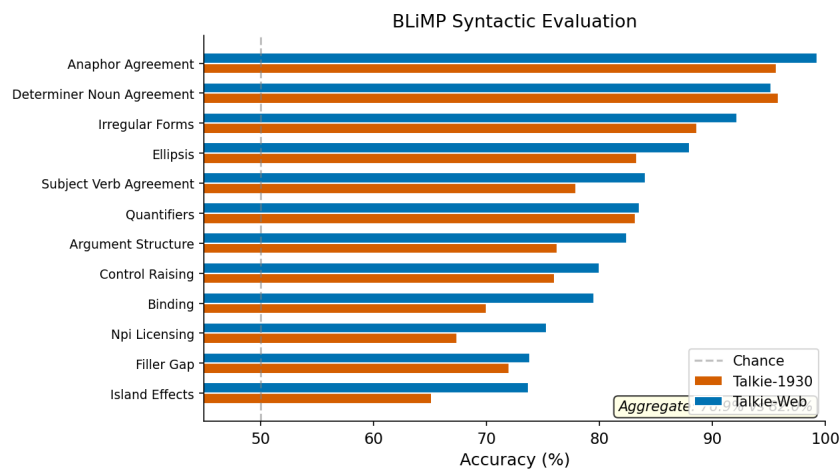


Figure 4: BLiMP accuracy by phenomenon (sorted by TALKIE-WEB). TALKIE-1930 (orange) tracks TALKIE-WEB (blue) across all 12 phenomena. Dashed line: chance (50%).

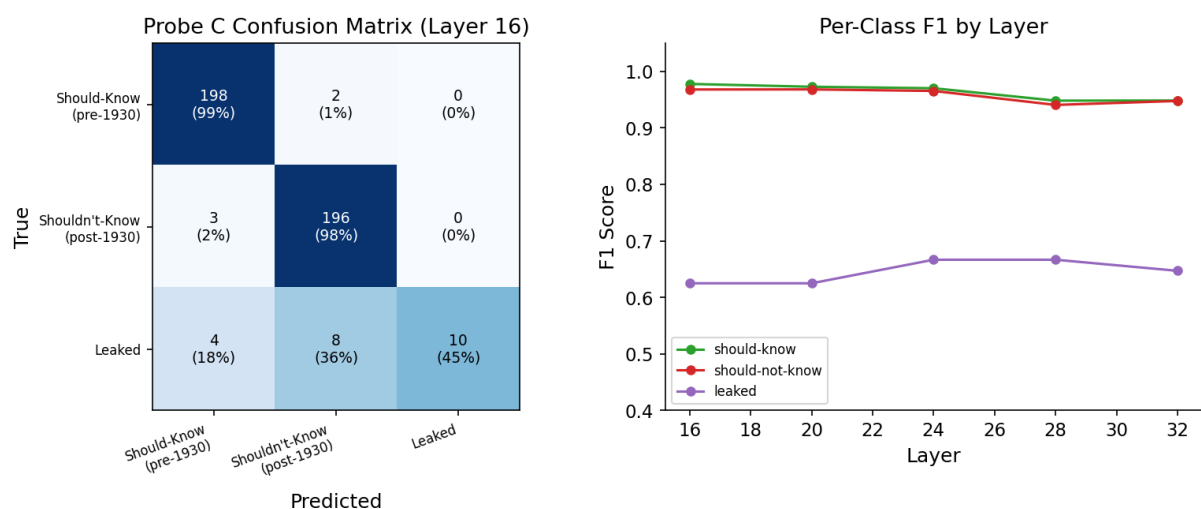


Figure 5: Probe C (TALKIE-1930). Left: confusion matrix at layer 16; should-know and should-not-know are cleanly separated, leaked items mostly confused with should-not-know. Right: per-class F1 by layer; the leaked class stays well below the others.

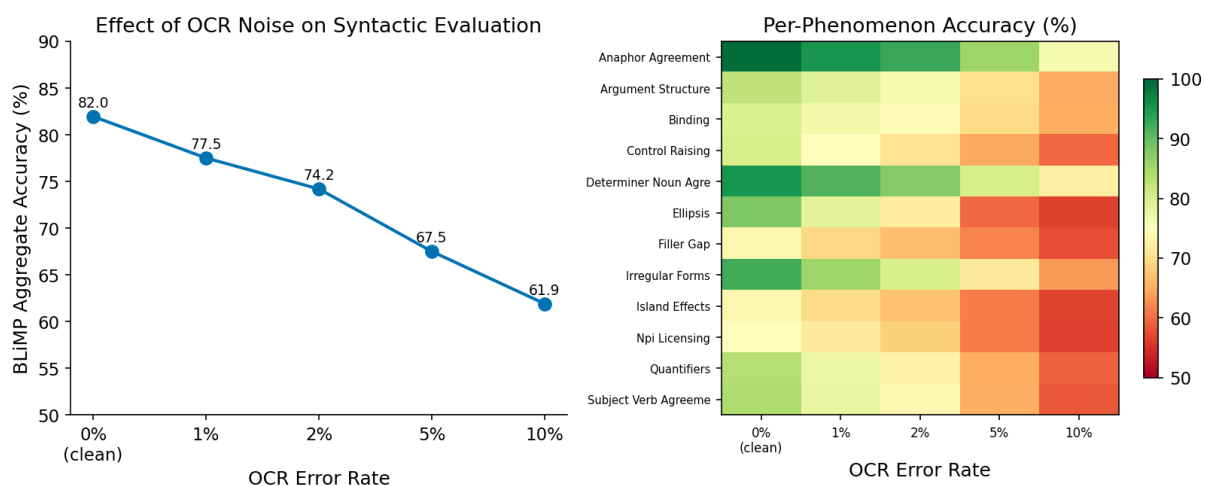


Figure 6: OCR noise ablation on TALKIE-WEB. Left: aggregate BLiMP accuracy degrades monotonically. Right: per-phenomenon accuracy; every phenomenon drops and none is preserved or improved.