

# Time Travel Engine: Ordered Temporal Geometry Enables Era-Level Historical Steering in Large Language Models

Jingmin An<sup>1</sup>, Wei Liu<sup>2</sup>, Qian Wang<sup>1,†</sup>, Fang Fang<sup>1,†</sup>

<sup>1</sup>Peking University, <sup>2</sup>Zhejiang University  
anjm@stu.pku.edu.cn, {wangqianpsy, ffang}@pku.edu.cn

## Abstract

Large language models compress centuries of discourse into fixed parameters, yet how they organize historical variation and whether this organization shapes generation remain unclear. We introduce the *Time Travel Engine* (TTE), a framework that compares geometric accounts of era-conditioned residual states and tests their behavioral accessibility under fixed prompts. Across six decoder language models in Chinese and English, broad eras repeatedly form ordered trajectories, although their exact spacing varies across model–language settings. Steering along these trajectories shifts both historical register and access to period-compatible knowledge, while adjacent anchors support local interpolation. Epistemic effects partly survive style removal, linking temporal steering to knowledge access beyond surface register. Across Chinese and English, partially alignable trajectories reveal corresponding chronological relations whose geometric realization remains language-specific. Together, ordered temporal geometry offers a testable account of how static language models organize diachronic variation, linking internal structure to historical register and access to period-compatible knowledge. Code and datasets: <https://github.com/LilTiger/Time-Travel-Engine>.

## 1 Introduction

*“Time is the substance I am made of. Time is a river which sweeps me along, but I am the river; it is a tiger which destroys me, but I am the tiger; it is a fire which consumes me, but I am the fire.”*

— *Borges* (1964, p. 220)

Time organizes change: in spacetime it serves as a coordinate, while Kant (1908) treats it as an a priori form through which experience is ordered. LLMs face an analogous compression problem. Their parameters absorb corpora spanning

centuries, but the resulting model state is static at inference time. We examine whether temporal information is represented, how diachronic variation is organized, and whether it can be causally controlled.

Recent work suggests that time is encoded inside LLMs. Gurnee and Tegmark (2023) identify linear temporal representations and “time neurons,” Papadopoulos et al. (2024) observe forward/backward prediction asymmetries, and Park et al. (2025) identify Temporal Heads for time-specific recall. Yet time-varying facts remain difficult: snapshot-trained models struggle with changing facts (Dhingra et al., 2022), and temporal misalignment can elicit outdated information (Zhang and Choi, 2023). Time is therefore represented, but it is not yet operationalized as a reliable control variable.

Representation engineering offers a route from analysis to intervention. Zou et al. (2023) place population-level representations at the center of interpretation and control, and contrastive activation addition steers behavior by adding residual-stream activation differences (Rimsky et al., 2024). However, steering vectors can be input-sensitive and brittle under prompt changes (Tan et al., 2024), while manifold steering improves over a single high-dimensional direction for overthinking in large reasoning models (Huang et al., 2025). For temporal control, this raises a concrete question: should historical time be modeled as discrete era anchors, or as an ordered trajectory that also supports intermediate states?

The question becomes harder cross-lingually. Shared structure can emerge in multilingual pre-training even without shared vocabulary (Conneau et al., 2020), but multilingual behavior can still depend on language-specific neurons (Tang et al., 2024), and cross-lingual consistency does not imply shared internal representations (Ifergan et al., 2024). Meanwhile, diachronic semantics exhibits measurable statistical regularities across time

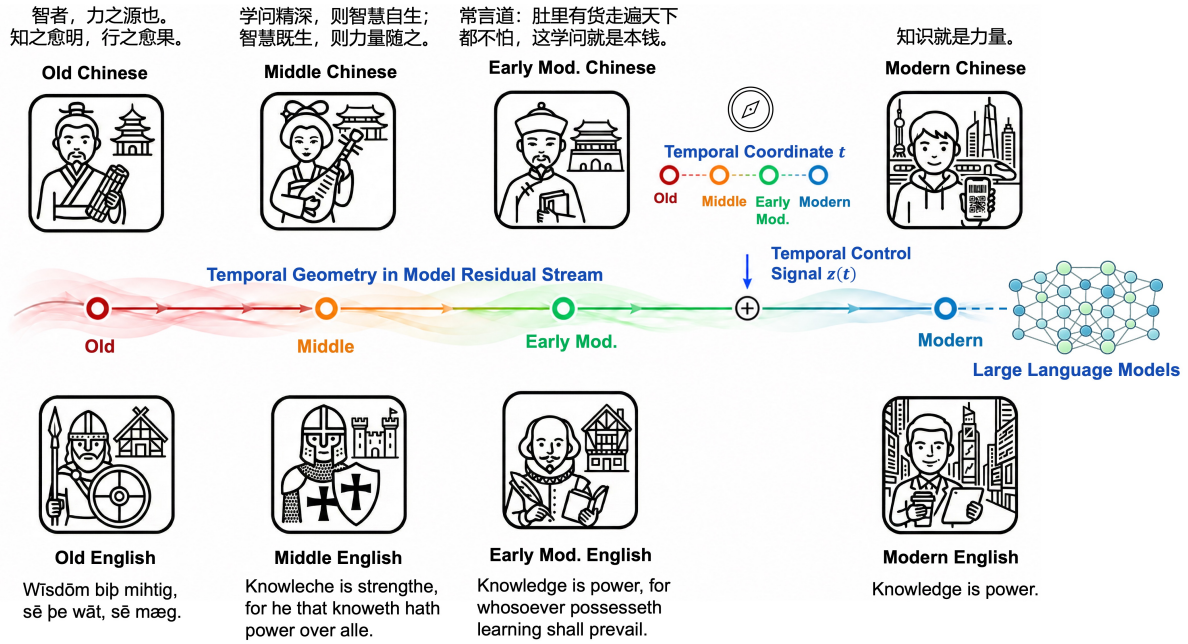


Figure 1: The Time Travel Engine (TTE). Parallel Chinese and English corpora across four broad eras define era-conditioned residual-stream anchors. A local coordinate  $t$  interpolates between adjacent learned anchors;  $z(t)$  applies the resulting control signal while the prompt remains fixed.

(Hamilton et al., 2016). These observations make cross-lingual transfer a natural probe of temporal steering: if temporal signals encode more than surface style, they should transfer across languages without requiring identical internal representations.

A central methodological concern is identifiability. Era is confounded in historical corpora with register, genre, topic, source, and cultural history; low-dimensional projections can also turn discrete clusters into apparent trajectories. A rigorous account must therefore distinguish whether era information is available, whether its relations are ordered beyond clustering, whether direct intervention changes behavior, and whether any effect persists under controls. Factual authority and historical understanding require independent evidence.

We introduce the *Time Travel Engine* (TTE, Figure 1) to make the temporal organization of model representations observable and controllable. Our findings establish three contributions:

- **Ordered temporal geometry.** Historical variation repeatedly forms broad-era order in the residual stream. Adjacent era anchors support local interpolation, while metric spacing varies across models and languages.
- **Behaviorally accessible historical states.** TTE steers both prose and verse toward a target period while prompts remain fixed, shift-

ing historical register and the epistemic boundary that governs which knowledge appears available. Part of this shift remains after estimated style removal.

- **Cross-lingual temporal organization.** Chinese and English trajectories are partially alignable, enabling cross-lingual transfer while revealing that temporal control remains coupled to linguistic form.

Together, these findings make era-correlated pre-training structure experimentally tractable and delineate the conditions under which it shapes generation. Calibrated dating, factual authority, and historical understanding require separate evidence.

## 2 Related Work

**Temporal Cognition in Language Models.** Early temporal NLP included deterministic temporal tagging and normalization systems such as SUTime (Chang and Manning, 2012). Recent work instead studies internal temporal structure in LMs: snapshot-trained models struggle with time-varying facts (Dhingra et al., 2022); LLMs learn linear temporal representations and identifiable “time neurons” (Gurnee and Tegmark, 2023); specific attention heads support recall of time-specific knowledge beyond simple numerical cues (Park et al.,

2025); and similarity judgments reveal Weber–Fechner-like temporal compression around a subjective reference point (Li et al., 2025).

**Representation Engineering (RepE).** Text control has progressed from training-time attribute control (Hu et al., 2017) toward representation-level analysis and intervention (Zou et al., 2023). Function Vectors show that a small set of attention heads can transport compact task representations (Todd et al., 2023), while Activation Addition and contrastive variants steer behavior through activation differences without weight updates (Turner et al., 2023; Rimskey et al., 2024). Subsequent work studies steering reliability (Tan et al., 2024), proposes unified evaluation principles (Im and Li, 2025), recasts demonstrations as latent in-context vectors (Liu et al., 2023), and derives theoretically grounded affine steering functions (Singh et al., 2024).

**Cross-Lingual Alignment and Representation Sharing.** One line of cross-lingual representation alignment studies unsupervised alignment of monolingual embedding spaces (Conneau et al., 2017). Yet multilingual capability can depend on language-specific neurons (Tang et al., 2024), intermediate processing may pass through a concept space closer to English (Wendler et al., 2024), and answer consistency can coexist with limited representation sharing (Ifergan et al., 2024). Consistently, LLMs may rely on dissimilar language-specific representations rather than a shared semantic space, although steering toward shared space improves transfer from, and consistency with, English (Lim et al., 2025). Related work also uses neuron activation modulation for style transfer with content consistency (Kong et al., 2025). We therefore test ordered broad-era structure, local interpolation between adjacent anchors, and partial cross-lingual trajectory alignment.

### 3 Methods

TTE follows an inferential sequence: it estimates era-conditioned residual directions under fixed prompt content, compares competing geometric accounts in activation space, intervenes on the recovered directions, and tests the resulting register and epistemic effects against surface, source, text-evidence, and cross-lingual alternatives.

#### 3.1 Models

We use Qwen2.5-14B-Base for the main-text analyses. We also analyze five additional decoder

| Period              | English      | Chinese               |
|---------------------|--------------|-----------------------|
| <i>Old</i>          | 450–1150     | 12th c. BC–3rd c. AD  |
| <i>Middle</i>       | 1150–1500    | 4th–12th c. AD        |
| <i>Early Modern</i> | 1500–1700    | 13th–early 20th c.    |
| <i>Modern</i>       | 1700–present | early 20th c.–present |

Table 1: Chronological periodization for English and Chinese used in our TTE framework.

models; their full names and results appear in Appendix A.2, with specifications in Table 30.

#### 3.2 Datasets

Our experiments use two main data resources: diachronic steering corpora and evaluation benchmarks. A separate contrastive style dataset for disentanglement is introduced in Section 3.5.

**Diachronic Steering Corpora.** To extract temporal features, we compiled English and Chinese diachronic corpora spanning four eras: *Old*, *Middle*, *Early Modern*, and *Modern*, following broad periodizations used in English and Chinese historical linguistics (van Gelderen, 2014; Dong, 2020) (Table 1). After cleaning, English prose contains 167K–182K tokens and verse 69K–73K tokens per era; Chinese prose contains 66K–78K tokens and verse 31K–37K tokens per era. Sources include the York-Toronto-Helsinki Parsed Corpus of Old English prose (YCOE) (Taylor, 2007), the Penn Parsed Corpora of Historical English (PPCHE) (Kroch, 2020), and the Chinese Text Project. The historical sources are nonparallel across topics and cultural periods, so source and topic variation remain part of the corpus-derived estimator.

**Evaluation Benchmarks.** To evaluate TTE beyond surface style transfer, we use three benchmarks targeting complementary aspects of temporal alignment: the Epistemic Cutoff Dataset (adherence to era-specific knowledge boundaries), the Causal Remodeling Dataset (historically plausible reasoning under period constraints), and the Mismatch Entanglement Dataset (style–cognition disentanglement under modern–historical mismatch). Appendix A.16 details their construction and gives representative evaluation queries.

#### 3.3 Characterizing Temporal Geometry

We first ask whether diachronic variation in the residual stream forms ordered temporal structure, and whether that structure is stronger than simpler clustered or stepwise accounts. Figure 2 motivates

this question visually, and formal tests are reported in Section 4.2. We begin by identifying a small set of era anchors.

Let  $\mathcal{M}$  be an LLM with  $L$  layers, and let  $\mathbf{h}_l(x) \in \mathbb{R}^d$  denote the hidden state at layer  $l$  for input  $x$ . We consider historical target eras  $\mathcal{T} = \{t_1, t_2, t_3\}$ , corresponding to *Old*, *Middle*, and *Early Modern*, with *Modern* as the reference era. We write  $\bar{\mathcal{T}} = \{\text{Old}, \text{Middle}, \text{Early Modern}, \text{Modern}\}$  for the complete anchor set. For interpolation experiments, anchors are placed at integer coordinates  $t_{\text{Old}} = 0$ ,  $t_{\text{Middle}} = 1$ ,  $t_{\text{E.Mod}} = 2$ , and  $t_{\text{Modern}} = 3$ , so that intermediate coordinates  $t = 0.5$  and  $t = 1.5$  correspond to local interpolations between adjacent anchor eras.

**Method I: Contrastive Activation Addition (CAA).** To identify the anchor of each target era, we adopt Contrastive Activation Addition (CAA) (Rimsky et al., 2024). For era  $t$ , we compare activations from self-generated responses to a shared prompt set  $\mathcal{P}$  under two *Era Charters* (Appendix A.14): the era-specific charter  $\mathcal{C}_t$  and the modern charter  $\mathcal{C}_0$ . Each charter is a system prompt that specifies the temporal persona of the model. Holding  $\mathcal{P}$  fixed isolates era-conditioned variation under identical prompt content:

$$\mathbf{v}_{\text{caa}}^{(l,t)} = \mathbb{E}_{x \sim \text{Gen}(\mathcal{P}|\mathcal{C}_t)}[\mathbf{h}_l(x)] - \mathbb{E}_{x \sim \text{Gen}(\mathcal{P}|\mathcal{C}_0)}[\mathbf{h}_l(x)] \quad (1)$$

For each accepted generation, we average at most the first 128 final-answer residual states and exclude prompt and Qwen3 reasoning states. For non-Thinking Old and Middle conditions, we truncate at the last stable six-token checkpoint if a style classifier favors Modern over the target era by more than .02 at three consecutive checkpoints; Early Modern and Qwen3 use the full available window, up to 128 tokens. Sample vectors are averaged within target and Modern conditions before differencing, yielding a synthetic era-conditioned anchor that includes the charter’s persona and register computation.

**Method II: Ensemble CAA (EnsCAA).** CAA controls prompt content, but its charter-conditioned generations still carry persona and register and sample less variation than the historical corpora. EnsCAA therefore combines the synthetic anchor with a corpus-derived centroid. We compute a parallel set of historical-corpus centroids, denoted

as  $\mathbf{v}_{\text{real}}^{(l,t)}$ , by averaging activations across the Diachronic Steering Corpora (Section 3.2). The ensemble vector  $\mathbf{v}_{\text{ens}}^{(l,t)}$  is formulated as a convex combination of the synthetic and corpus-derived estimators:

$$\mathbf{v}_{\text{ens}}^{(l,t)} = \alpha \cdot \mathbf{v}_{\text{caa}}^{(l,t)} + (1 - \alpha) \cdot \mathbf{v}_{\text{real}}^{(l,t)} \quad (2)$$

For  $\mathbf{v}_{\text{real}}$ , each document is split into nonoverlapping chunks of at most 128 tokens; fragments of at most 16 tokens are discarded, non-padding states are averaged within each chunk, and retained chunks are averaged within era. The estimator is therefore chunk-weighted. We use  $\alpha = .6$  for EnsCAA.

**Method III: Chronological Manifold Projection (CMP).** Method I yields discrete era anchors. Chronological Manifold Projection (CMP) tests local interpolation between adjacent anchors. We treat the discrete synthetic centroids from Method I,  $\mathcal{V} = \{\mathbf{v}_{\text{caa}}^{(l,t)} \mid t \in \bar{\mathcal{T}}\}$ , as anchor points defining a high-dimensional trajectory. We first identify a principal temporal subspace  $\mathbf{U}_l \in \mathbb{R}^{d \times k}$  via PCA. Within this latent geometry, we fit a polynomial spline  $\mathcal{S}(\cdot)$  to the projected coordinates, yielding

$$\mathbf{v}_{\text{cmp}}^{(l)}(t) = \mu_l + \mathbf{U}_l \mathcal{S}(t; \mathcal{V}) \quad (3)$$

This formulation provides a local interpolation operator between adjacent anchors; its behavior is evaluated empirically through generated outputs and raw-space measurements.

**Method IV: Ensemble Manifold Projection (EnsCMP).** We apply CMP to the ensemble centroids  $\{\mathbf{v}_{\text{ens}}^{(l,t)}\}$  with  $\alpha = .7$ . The resulting  $\mathbf{v}_{\text{ens-cmp}}^{(l,t)}$  preserves anchor order while incorporating observed historical text; corpus topic and source distributions remain part of this estimator.

**Geometric Hypothesis Tests.** Era anchors yield era-conditioned hidden states whose pairwise distance matrix  $D_{\text{rep}}$  is then compared to chronological distance  $D_{\text{time}}$  between anchors. We distinguish three nested accounts: (i) a *temporal-distance* account in which  $D_{\text{rep}}$  grows monotonically with  $D_{\text{time}}$ ; (ii) a *strict-cluster* account that allows only a within-era versus across-era contrast; and (iii) an *ordered-stepwise* account that requires monotone era ordering without metric continuity. We test (i) via the Mantel correlation  $r(D_{\text{rep}}, D_{\text{time}})$  and compare model fits via  $\Delta\text{BIC}$ , the difference

in the Bayesian Information Criterion between accounts (positive values favor temporal distance over each alternative). Held-out decade projection and unseen intermediate coordinates evaluate *local* interpolation between adjacent anchors.

### 3.4 Temporal Intervention Mechanism

We modulate temporal signals with a context-relative dose  $d_l = \lambda \|\mathbf{h}_l\|_2$  along a unit-normalized era direction. This rule is L2-sensitive: an unusually large activation coordinate increases the total norm and therefore the applied dose. The modified state is

$$\tilde{\mathbf{h}}_l = \mathbf{h}_l + \lambda \cdot \|\mathbf{h}_l\|_2 \cdot \frac{\mathbf{v}^{(l)}}{\|\mathbf{v}^{(l)}\|_2} \quad (4)$$

We use  $\lambda \in [.05, .15]$  and the unclipped norm in the reported experiments; Appendix A.1.5 tests 95th- and 99th-percentile activation-norm caps (p95/p99). For each model–language–era–method setting, Log Probability Dominance Ratio (LPDR) is the in-period minus future-entity mean log probability, and layers are ranked by  $J_l = .5z(F_l A_l) + .5z(\Delta \text{LPDR}_l - .25P_l)$ . Here  $F_l$  is Fisher separation,  $A_l$  is held-out probe accuracy,  $\Delta \text{LPDR}$  is the intervention gain in this ratio, and  $P_l = \max(\Delta \text{loss} - \tau, 0)$  is the excess-loss penalty with  $\tau = .03$ . The five highest-scoring available layers are fixed before final downstream evaluation; Appendix A.1.1 details the selection protocol and its fixed components.

### 3.5 Time-Style Disentanglement

A central question is whether temporal steering acts mainly through diachronic surface register or through historically conditioned knowledge. To probe this relation, we construct a Contrastive Style Dataset of semantically aligned expressions across historical registers, with representative examples given in Appendix A.15. From the activation differences between these stylistic pairs, we estimate a dominant linear style subspace  $\mathbf{U}_{style}$  using PCA and retain the top- $k$  components, with  $k = 5$  in our experiments.

Given a raw time vector  $\mathbf{v}_{time}$ , we attenuate its leading stylistic component by removing the projection onto this subspace:

$$\mathbf{v}_{cog} = \mathbf{v}_{time} - \text{proj}_{\mathbf{U}_{style}}(\mathbf{v}_{time}) \quad (5)$$

Here,  $\mathbf{v}_{cog}$  denotes the style-residual direction obtained after removing the estimated linear style

subspace. Its interpretation is limited to that operational residual. Steering with  $\mathbf{v}_{cog}$  tests whether temporally appropriate knowledge and historical boundary control persist after this removal.

To test whether residual entanglement is an artifact of linear PCA, we also train a nonlinear autoencoder (AE), implemented as a multilayer perceptron (MLP),  $f_\theta = D \circ E$  on the same style-difference activations for 100 epochs. The autoencoder uses bottleneck  $k = 5$  and is fit on the top five layers identified by the same layer-selection procedure used for steering. Given  $\mathbf{v}_{time}$ , we initialize a latent code from  $E(\mathbf{v}_{time})$  and optimize it for 200 steps to find the closest decoded style point  $D(\mathbf{z}^*)$  on the learned nonlinear manifold. The nonlinear style residual is then  $\mathbf{v}_{cog}^{AE} = \mathbf{v}_{time} - D(\mathbf{z}^*)$ . Linear and nonlinear residuals are evaluated with the same FLR/PR protocol.

### 3.6 Cross-Lingual Temporal Transfer

To test whether a temporal direction learned in one language can steer another language, we consider two settings. Under *direct transfer*, the source-language direction  $\mathbf{v}_{src}^{(t)}$  is applied to target-language prompts without coordinate alignment. Under *manifold-aligned transfer*, we first estimate an orthogonal Procrustes map  $\mathbf{R}$  from the source temporal trajectory to the target:

$$\min_{\mathbf{R} \in O(d)} \sum_t \|\mathbf{v}_{tgt}^{(t)} - \mathbf{R} \cdot \mathbf{v}_{src}^{(t)}\|_F^2 \quad (6)$$

We then apply  $\mathbf{R}\mathbf{v}_{src}^{(t)}$  to target-language prompts. The map is estimated from four ordinal era anchors. Direct transfer tests reuse of a source-language direction, whereas manifold-aligned transfer tests correspondence between era-level trajectories. Absolute calendar equivalence, concept translation, and cultural equivalence require separate evidence.

**Epistemic Integrity.** To assess whether generation respects the epistemic boundary of target era  $t$ , let  $E \equiv E_{gen}$  denote the set of generated entities, and let  $\tau(e)$  denote the era assigned to entity  $e$  by an external knowledge base. We define

$$\begin{aligned} \text{FLR}(t) &= \frac{1}{|E|} \sum_{e \in E} \mathbb{I}[\tau(e) > t], \\ \text{PR}(t) &= \frac{1}{|E|} \sum_{e \in E} \mathbb{I}[\tau(e) \leq t]. \end{aligned} \quad (7)$$

Here, FLR measures future leakage beyond era  $t$ , whereas PR measures chronological precision within era  $t$ .

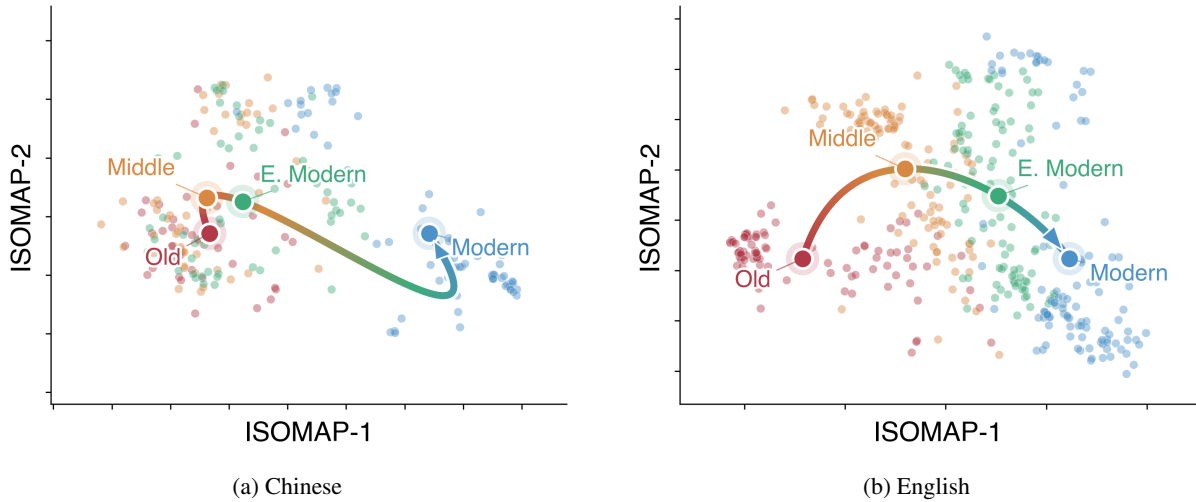


Figure 2: Visualization of four-era temporal structure in Qwen2.5-14B-Base. Isomap projections show ordered era-conditioned trajectories in Chinese and English; formal competing-hypothesis tests quantify the corresponding raw-space structure in Figure 3.

## 4 Results

A temporal representation becomes behaviorally meaningful when its chronological organization can be linked to systematic changes in generation. We first examine the historical form of steered outputs, then establish whether their underlying activations are ordered beyond clustering and whether this organization shapes epistemic access, persists after estimated style removal, and transfers across languages. Main-text analyses use Qwen2.5-14B-Base; model-wise results, alternative estimators, and controls appear in Appendices A.1–A.6.

### 4.1 Qualitative Diachronic Steering

We first ask whether temporal vectors jointly alter lexical form, syntax, and period-specific assumptions. Tables 2 and 3 compare the unsteered modern baseline with outputs modulated by *Old*, *Middle*, and *Early Modern* signals. The *Old Chinese* vector induces Classical vocabulary and syntax (“夫”, “之”), while the *Early Modern English* vector induces Shakespearean syntax and orthography (“thou”, “thyselfe”) together with period-specific assumptions about how a good life should be lived.

### 4.2 Broad-Era Order Recurs beyond Strict Clustering

We next test whether the visual geometry in Figure 2 reflects chronological distance or a weaker alternative. We compare three accounts in raw activation space using Mantel correlation and  $\Delta\text{BIC}$  model selection. The temporal-

distance account predicts that representation distance grows monotonically with chronological distance; the strict-cluster account treats all cross-era distances as equivalent; and the ordered-stepwise account preserves era order without requiring metric continuity. Raw activations strongly favor temporal distance over strict clusters in Chinese ( $r=0.341$ ,  $\Delta\text{BIC}=172.0$ ) and English ( $r=0.602$ ,  $\Delta\text{BIC}=564.2$ ). Against the harder ordered-stepwise baseline (Figure 3), the temporal-distance account gives the better fit in Chinese ( $\Delta\text{BIC}=14.9$ ), whereas the ordered-stepwise account fits slightly better in English ( $-4.3$ ). Together, the two comparisons identify ordered temporal geometry as the shared structure: Chinese supports a metric account, while English behaves more like a coarse sequence of ordered intervals under the same criterion. Across all six evaluated models, Mantel correlations are positive in 12/12 model–language pairs; temporal distance outperforms strict clusters in 11/12 and the ordered-stepwise account in 5/12 (Appendix A.4). The recurring property is order, while metric spacing remains model- and language-dependent.

### 4.3 Finer-Scale Order Is Detectable but Uneven

Held-out decades provide a stricter test because the 1900s–1990s activations are never used to fit the four-era reference geometry. Figure 4 shows fine ordering in English ( $\rho=0.745$ ,  $p=0.013$ ,  $N=10$ ) and a weaker positive trend in Chinese ( $\rho=0.418$ ,  $p=0.229$ ). Unseen decades can therefore preserve

**PROMPT**

人应该如何度过一生？

*How should one live a good life?***Baseline (Modern State)**

人生是一个复杂而独特的过程，每个人都有自己独特的经历和价值观...

*Life is a complex and unique journey, where everyone possesses their own distinct experiences and values...***Signal: Old Chinese**

有二事为大，一曰孝，二曰敬。《礼记》云：夫孝，德之本也...

*Two matters are paramount: Filial Piety and Reverence. The Book of Rites states: "Now, filial piety is the root of all virtue..."***Signal: Middle Chinese**

惟有三事可为：一曰孝，二曰悌，三曰敬。此三者乃为人之本也...

*There are but three pursuits worthy of action: Filial Piety, Fraternal Duty, and Reverence. These three constitute the very foundation of being human.***Signal: Early Modern Chinese**

人生苦短，如何度日方好？《菜根谭》曰：天欲祸人，必先以微福骄之...

*Life is bitterly short; how then should one pass one's days? The Vegetable Root Discourse says: "When Heaven intends to bring calamity upon a man, it first puffs him up with trifling blessings..."*

Table 2: Chinese diachronic register modulation via TTE on Qwen2.5-14B-Base; gray text gives English translations.

finer order within the broad-era geometry, with language-dependent resolution. Appendix A.1.2 reports the full six-model estimates, heterogeneity, and multiplicity-adjusted tests.

Unseen-coordinate generations complement the activation-space decade test by asking whether interpolated coordinates produce mixed era evidence in output. We compute era probabilities using a held-out character  $n$ -gram classifier over the four anchor labels and interpret the probabilities as soft

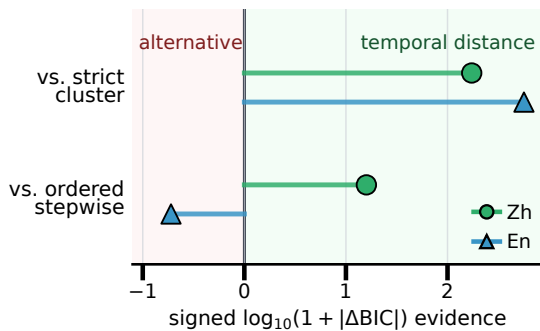


Figure 3: Competing-hypothesis tests for temporal geometry on Qwen2.5-14B-Base; positive signed log-BIC evidence favors temporal distance over strict-cluster and ordered-stepwise alternatives.

**PROMPT**

How should one live a good life?

**Baseline (Modern State)**

This is the question that has been asked by philosophers for thousands of years. The answer to...

**Signal: Old English**

Se þe wille gepyldlice lifian sceal healdan godne gewunum, swa swa se gecweneð on his gereorde...

*He who wishes to live steadfastly must maintain good customs, just as the speaker declared in his speech...***Signal: Middle English**

In the olde tymes of yore, whan lewys and lyf were grete, men soughte to fynde that which maketh a good lyfe..

*In the olden times of yore, when laws and life were great, men sought to find that which maketh for a good life...***Signal: Early Modern English**

Thou shalt know thyselfe first, then seeke vertue, befiege it with thyne endeaours. To liue well is to liue vertuouslie.

*First, know thyself; then seek virtue, and pursue it with all thy might. To live well is to live virtuously.*

Table 3: English diachronic register modulation via TTE on Qwen2.5-14B-Base.

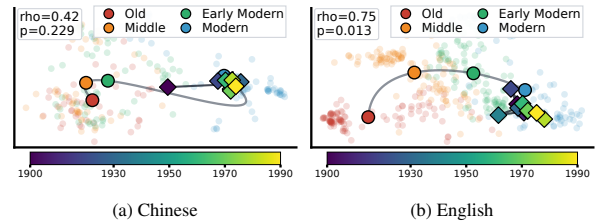


Figure 4: Held-out decade projection for Qwen2.5-14B-Base: 1900s–1990s activations are projected into the fixed four-era reference geometry without refitting.

evidence, not as exact dates. Table 4 shifts from Old- to Middle-dominant in English and more gradually in Chinese, matching local movement between neighboring anchors. Across architectures, raw-space order can remain detectable even when reasoning-time decoding weakens generation-level control (Appendix A.6). This dissociation separates representational order from behavioral accessibility.

#### 4.4 Epistemic Integrity and Knowledge Boundaries

Temporal control must affect more than orthography. The Epistemic Cutoff Dataset jointly measures future leakage (FLR) and era-appropriate precision (PR) over philosophers, inventions, and tools; the unsteered baseline returns modern entities for every era. On Qwen2.5-14B-Base, En-SCMP versus CAA lowers FLR (0.338→0.260) and raises PR (0.362→0.607) across *Old*, *Middle*, and *Early Modern* (Figure 5; Appendix A.2.1).

| Lang. | Coord.    | Old   | Middle | E. Mod. | Modern |
|-------|-----------|-------|--------|---------|--------|
| Zh    | $t = 0.5$ | 0.313 | 0.265  | 0.244   | 0.178  |
| Zh    | $t = 1.5$ | 0.290 | 0.257  | 0.238   | 0.216  |
| En    | $t = 0.5$ | 0.341 | 0.239  | 0.174   | 0.246  |
| En    | $t = 1.5$ | 0.223 | 0.343  | 0.209   | 0.225  |

Table 4: Unseen intermediate soft era probabilities in Qwen2.5-14B-Base.

This joint shift reduces the abstention explanation because high PR requires era-admissible entities; blind-rubric and randomized-network controls assess coherence separately (Appendices A.1.5 and A.8). Causal Remodeling shows a parallel change: *Old*-era steering replaces electric lighting with oil lamps and rush candles. The intervention therefore changes era-compatible entities alongside register, within pretrained knowledge.

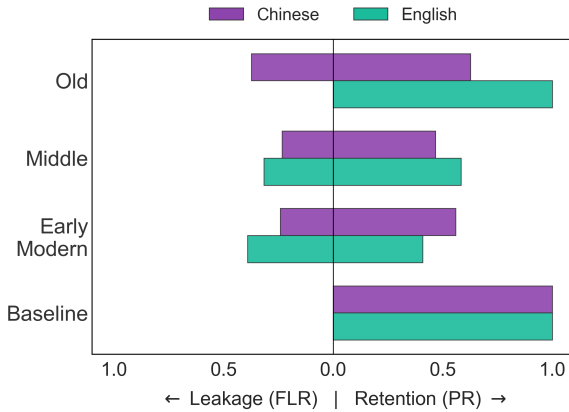


Figure 5: Epistemic boundary test for Qwen2.5-14B-Base: EnsCMP lowers future leakage (FLR) while improving era-appropriate retention (PR).

#### 4.5 Epistemic Effects Partly Survive Style Removal

Style removal asks whether the epistemic effect is a by-product of archaic register. We compare linear PCA with nonlinear MLP-autoencoder residuals on the same style-difference activations (Section 3.5). Both leave temporal control partially intact; model-wise AE results are in Table 13. Figure 6 shows the linear PCA style-removal trajectory in (PR, FLR) space. Residual vectors continue to move FLR and PR after the leading style subspace is removed, while weaker behavior on older targets shows that historical form carries part of the signal. The temporal vector therefore combines a persistent epistemic component with style-dependent structure.

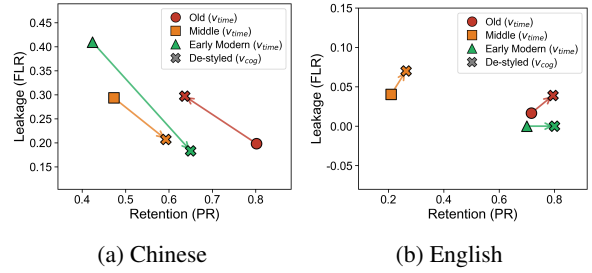


Figure 6: Time-style entanglement on Qwen2.5-14B-Base after style removal ( $\mathbf{v}_{time} \rightarrow \mathbf{v}_{cog}$ ): arrows show the metric shift in (PR, FLR) space.

#### 4.6 Temporal Trajectories Are Partially Alignable across Languages

Cross-lingual transfer provides a stringent test of what is shared across languages. The question is whether chronological relations among era-conditioned states survive a change of linguistic coordinate system, even when the directions that realize those relations differ. Direct transfer asks whether source and target trajectories are already commensurate. Manifold-aligned transfer maps the four source-language anchors onto the target trajectory before steering (Section 3.6). Comparing them distinguishes reuse of a common direction from correspondence between ordered structures.

For Qwen2.5-14B-Base, manifold-aligned transfer is most effective with EnsCMP: relative to manifold-aligned CAA, FLR decreases from .701 to .196 and PR rises from .146 to .371 (Appendix A.2.3). This paired shift shows that relations among multiple era anchors can make temporal organization accessible across languages. Across all six models, however, the estimator that best balances FLR and PR differs by model. The cross-lingual regularity is therefore relational and partial: chronological organization can correspond across languages, while its effective realization remains model-dependent.

The generations reveal how this correspondence is realized. For the same winter-storm request, native English directions preserve English while moving the register across periods. Chinese directions evoke period-associated Chinese forms, including Classical Chinese under *Old* (Table 5), and the converse transfer appears for a Chinese request in Table 29 in the Appendix. This bidirectional pattern shows that era-related organization is accessible across languages; the accompanying language switches show that it remains coupled to linguistic form. Together, the quantitative transfer and bidi-

| <b>PROMPT</b>                                               |                                                                                                                                                                                   |                                                                                                                                                                            |
|-------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Describe a winter storm in verse.                           |                                                                                                                                                                                   |                                                                                                                                                                            |
| <b>Baseline (Modern State)</b>                              | <b>Native Direction (EN→EN)</b>                                                                                                                                                   | <b>Cross-Lingual Direction (ZH→EN)</b>                                                                                                                                     |
| <b>Direction: Old</b>                                       |                                                                                                                                                                                   |                                                                                                                                                                            |
| The sky was dark and heavy, A blanket of grey, so dreary... | Hwæt! Þisses wintres storme is geworhtan þurh miclan þrymm þeoda gewyrdum...<br><i>Listen! This winter storm has been wrought by great might, through the fates of peoples...</i> | 惟天有雪，其寒若霜。风兮何起，云兮何茫...<br><i>Heaven alone holds the snow; its cold is like frost. Whence rises the wind; how vast are the clouds...</i>                                    |
| <b>Direction: Middle</b>                                    |                                                                                                                                                                                   |                                                                                                                                                                            |
| The sky was dark and heavy, A blanket of grey, so dreary... | þe wynd doth blowe wyth fyer and might, þe snow falleþ downe styffe and bryght...<br><i>The wind blows with fire and might, The snow falls down stiff and bright...</i>           | 惟馮雪皓然，白者皆素焉。寒月霜气冽，清风松韵传...<br><i>Only snow gleams; all is white. Frost under the cold moon bites; clear wind carries the pines' resonance...</i>                           |
| <b>Direction: Early Modern</b>                              |                                                                                                                                                                                   |                                                                                                                                                                            |
| The sky was dark and heavy, A blanket of grey, so dreary... | A strong wisterne winde doth blow, And with his strength he sendes forth snow...<br><i>A strong western wind does blow, And with his strength he sends forth snow...</i>          | 南枝梅蕊，北窗竹影，西阁茶烟。雪风儿六花作团，云月儿一玉相联...<br><i>Plum buds south, bamboo shadows north, tea smoke west. Snowy winds gather six-petaled flakes; clouds and moon unite as jade...</i> |

Table 5: Qualitative cross-lingual register modulation on Qwen2.5-14B-Base using an English query. The baseline is compared with native-language and cross-lingual temporal directions.

rectional generations identify partially alignable temporal organization with language-specific realizations. Target-language preservation and cultural equivalence remain separate objectives.

#### 4.7 Era Directions Correspond to Held-Out Historical Distributions

The recovered directions also correspond to held-out historical language distributions. We estimate directions on 80% of each diachronic corpus and evaluate perplexity (PPL) on the held-out 20%, column-normalizing model–language matrices before averaging across the six models. Same-era historical PPL is lower than off-diagonal PPL in Chinese (.476 vs. .652) and English (.367 vs. .678; Figure 7 in the Appendix). This same-era advantage connects the recovered directions to held-out historical distributions, while the lower unsteered PPL marks absolute language-model fit as a sepa-

rate objective.

## 5 Conclusion

Historical variation leaves ordered, behaviorally accessible geometry in decoder residual streams. TTE tests this structure through geometric comparison, held-out chronology, and intervention under fixed prompts. Broad eras recur in order, adjacent anchors support local interpolation, and steering changes historical register and epistemic boundaries; part of the latter survives style removal. Chinese and English trajectories are partially alignable, revealing cross-lingual temporal organization coupled to linguistic form. Together, they connect diachronic variation in representation and generation. Geometry and controllability vary across models, languages, eras, and decoding. Ordered temporal geometry thus grounds the study of how static models organize changing language and knowledge.

## 6 Limitations

TTE inherits the coverage and biases of both pretraining data and the historical reference corpora. Same-content, marker-ablation, and source-controlled corpus comparisons constrain several surface explanations, yet uneven preservation, genre imbalance, and culturally specific concepts can remain encoded in the recovered directions. Moreover, temporal register and epistemic access are only partly separable: style removal preserves part of the signal, so era directions need not isolate historical knowledge from linguistic form. The intervention also depends on activation scale; norm-based dosing is stable in the evaluated settings, while longer contexts and other decoding regimes require further sensitivity analysis. The intervention establishes that the selected residual directions can alter outputs under the evaluated conditions. Whether natural generation relies on the same pathway, and whether the resulting content is historically accurate, require independent evidence.

**Potential Risks.** Chronological steering may expose behaviors associated with eras that predate modern safety norms, reducing the effectiveness of safety alignment calibrated on contemporary data. It can also compress diverse historical communities into a small set of control directions, amplifying biases or stereotypes and encouraging generated accounts that depart from factual reconstruction. These risks motivate explicit safety evaluation and source-grounded validation before any use beyond representational analysis.

## Acknowledgments

This study was supported by the National Science and Technology Innovation 2030 Major Program (2022ZD0204802), the National Natural Science Foundation of China (T2421004), and the National Key Research and Development Program of China (2025YFE0218200). Generative AI systems (GPT and Claude) were used for language editing, code assistance, and blinded rubric analysis.

## References

- Jorge Luis Borges. 1964. *Labyrinths: Selected Stories & Other Writings*. New Directions, New York. Edited and translated by Donald A. Yates and James E. Irby.
- Angel X. Chang and Christopher D. Manning. 2012. [SUTime: A library for recognizing and normalizing](#)

[time expressions](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 3735–3740, Istanbul, Turkey. European Language Resources Association (ELRA).

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

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

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.

Hongyuan Dong. 2020. *A history of the Chinese language*. Routledge.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Yao Huang, Huanran Chen, Shouwei Ruan, Yichi Zhang, Xingxing Wei, and Yinpeng Dong. 2025. Mitigating overthinking in large reasoning models via manifold steering. *arXiv preprint arXiv:2505.22411*.

- Maxim Ifergan, Leshem Choshen, Roei Aharoni, Idan Szpektor, and Omri Abend. 2024. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in llms. *arXiv preprint arXiv:2408.10646*.
- Shawn Im and Sharon Li. 2025. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*.
- Immanuel Kant. 1908. Critique of pure reason. In Benjamin Rand, editor, *Modern Classical Philosophers*, pages 376–456. Houghton Mifflin, Boston.
- Chaona Kong, Jianyi Liu, Yifan Tang, and Ru Zhang. 2025. Neuron activation modulation for text style transfer: Guiding large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7735–7747.
- Anthony Kroch. 2020. [Penn parsed corpora of historical english \(ldc2020t16\)](#). Web Download.
- Lingyu Li, Yang Yao, Yixu Wang, Chubo Li, Yan Teng, and Yingchun Wang. 2025. The other mind: How language models exhibit human temporal cognition. *arXiv preprint arXiv:2507.15851*.
- Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2025. Language-specific latent process hinders cross-lingual performance. *arXiv preprint arXiv:2505.13141*.
- Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2023. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.
- Vassilis Papadopoulos, Jérémie Wenger, and Clément Hongler. 2024. Arrows of time for large language models. *arXiv preprint arXiv:2401.17505*.
- Yein Park, Chanwoong Yoon, Jungwoo Park, Minbyul Jeong, and Jaewoo Kang. 2025. Does time have its place? temporal heads: Where language models recall time-specific information. *arXiv preprint arXiv:2502.14258*.
- Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.
- Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roei Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. 2024. Representation surgery: Theory and practice of affine steering. *arXiv preprint arXiv:2402.09631*.
- Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715.
- Ann Taylor. 2007. [The york-toronto-helsinki parsed corpus of old english prose](#). In *Creating and Digitizing Language Corpora*, pages 196–227. Palgrave Macmillan UK.
- Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.
- Elly van Gelderen. 2014. [A History of the English Language](#), second edition. John Benjamins Publishing Company.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Michael Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14213–14226.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

## A Appendix

Appendix A.1 reports extraction, layer selection, held-out chronology, confound controls, alternative mechanisms, blind rubric evaluation, and norm sensitivity. Appendix A.2 reports per-model scaling on world knowledge, style-removal disentanglement, and cross-lingual transfer. Appendix A.4–A.6 evaluates competing geometric hypotheses, decade projection, and unseen coordinates. Appendix A.7–A.10 gives architecture, randomization, sparse-feature, and prompt-conflict controls; Appendix A.11–A.16 documents examples, models, and datasets.

Tables use the following model abbreviations: Qwen-14B (Qwen2.5-14B-Base), Qwen-14B-Inst (Qwen2.5-14B-Instruct), Llama-8B-Inst (Llama-3.1-8B-Instruct), Gemma-9B-Inst (Gemma-2-9B-Instruct), Qwen3-30B (Qwen3-30B-A3B-Instruct), and Qwen3-4B-Think (Qwen3-4B-Thinking).

### A.1 Additional Controls and Reproducibility

The central interpretation requires era-directed effects to survive plausible alternatives: active roleplay, visible historical markers, source and topic differences, added text, parameter adaptation, incoherent output, and unusually large activation coordinates. The controls below isolate these alternatives one at a time. Unless stated otherwise, intervals are 95% confidence intervals.

#### A.1.1 Extraction, Layer Selection, and Intervention

Temporal directions are estimated from two complementary sources. For CAA, we average at most the first 128 answer-side residual states within each accepted final answer, exclude prompt and reasoning states, and subtract the Modern condition mean from the target-era mean. For the corpus-derived estimate, historical documents are divided into nonoverlapping chunks of at most 128 tokens; fragments shorter than 16 tokens are discarded, and retained chunks receive equal weight within an era. EnsCAA combines the two estimates with weight  $\alpha = .6$ .

CMP asks whether the four CAA anchors support local movement between adjacent eras: it fits a PCA subspace and polynomial spline through those anchors. EnsCMP applies the same construction to ensemble centroids using  $\alpha = .7$ . No interpolation is evaluated outside adjacent anchor intervals. For every model, language, target era, and method, the

five highest-scoring residual layers are fixed before downstream evaluation. The Modern condition is the reference anchor and therefore has no target steering layers. At inference, a unit-normalized era direction is applied with  $\lambda \in [.05, .15]$ ; the primary results use the unclipped hidden-state norm.

Layer selection balances three properties: separation between target and Modern states, generalization to held-out activations, and intervention gain without a large loss increase. The fixed objective in Section 3.4 is evaluated on an 80/20 activation split and 10 ontological items per language. The primary FLR/PR results use the separate Epistemic Cutoff Dataset, so the reported benchmark outcomes do not determine the selected layers. The ontological items also contribute to Causal Remodeling, and this separation claim therefore applies specifically to the primary Epistemic Cutoff evaluation.

#### A.1.2 Held-Out Finer Chronology

Each of ten 1900s–1990s decades is projected into a fixed four-era geometry without refitting. The random-effects pooled Spearman association is .566 ( $p < .001$ ,  $I^2 = 48.1\%$ ); the Chinese and English estimates are .686 and .416, respectively. Five of 12 model–language pairs survive Benjamini–Hochberg correction. The four anchors therefore recover finer chronological order in aggregate, with substantial variation across models and languages.

#### A.1.3 Same-Content and Visible-Marker Controls

A passive-content control asks whether temporal order remains when semantics are fixed and no era charter, free generation, or roleplay is involved. For each language, we teacher-force four era realizations of 200 items and hold out items in five folds. All 12 model–language combinations retain positive held-out order (permutation  $p \leq .008$ ). Separately, era-discriminative visible markers are identified from the training folds and removed before reclassification; only 92 of 9,600 generation labels change. Temporal order therefore persists under fixed semantics and after removal of a compact marker lexicon, although distributed register and source cues may remain.

#### A.1.4 Source-Controlled Centroids and External Baselines

To reduce source and topic imbalance, we pair historical passages within language, era, and genre by semantic similarity and length, then recompute

| Model          | Zh $\rho$ | Zh BH- $q$ | En $\rho$ | En BH- $q$ |
|----------------|-----------|------------|-----------|------------|
| Llama-8B-Inst  | .830      | .015       | -.212     | .668       |
| Qwen-14B       | .418      | .344       | .745      | .040       |
| Qwen-14B-Inst  | .903      | .004       | .018      | .960       |
| Qwen3-30B      | .721      | .045       | .139      | .765       |
| Qwen3-4B-Think | .248      | .652       | .818      | .015       |
| Gemma-9B-Inst  | .636      | .096       | .539      | .184       |

Table 6: Held-out decade ordering in the fixed four-era geometry ( $N = 10$  decades per model–language pair).

| Model          | Zh $\rho$ [95% CI] | En $\rho$ [95% CI] | Zh strict | En strict |
|----------------|--------------------|--------------------|-----------|-----------|
| Llama-8B-Inst  | .110 [.056, .163]  | .946 [.930, .960]  | .010      | .755      |
| Qwen-14B       | .366 [.330, .406]  | .973 [.962, .983]  | .035      | .875      |
| Qwen-14B-Inst  | .383 [.346, .420]  | .983 [.972, .992]  | .030      | .930      |
| Qwen3-30B      | .362 [.324, .402]  | .977 [.965, .987]  | .040      | .900      |
| Qwen3-4B-Think | .276 [.235, .317]  | .951 [.936, .965]  | .020      | .785      |
| Gemma-9B-Inst  | .210 [.165, .258]  | .987 [.978, .994]  | .015      | .945      |

Table 7: Five-fold temporal ordering with semantic content held fixed; strict is the fraction of items with fully monotonic four-era predictions.

| Model          | Labels | Changes | Rate   |
|----------------|--------|---------|--------|
| All            | 9,600  | 92      | .958%  |
| Llama-8B-Inst  | 1,600  | 20      | 1.250% |
| Qwen-14B       | 1,600  | 10      | .625%  |
| Qwen-14B-Inst  | 1,600  | 12      | .750%  |
| Qwen3-30B      | 1,600  | 25      | 1.563% |
| Qwen3-4B-Think | 1,600  | 14      | .875%  |
| Gemma-9B-Inst  | 1,600  | 11      | .688%  |

Table 8: Era-label changes after visible-marker removal.

the corpus-derived centroids. We use 1,536 such passage pairs and evaluate the resulting centroids on 180 queries per model. Their temporal log-odds (TLO) shifts over prompting remain positive for Qwen-14B and Qwen-14B-Inst, whereas the Qwen3-4B-Think estimate is uncertain. This control preserves the effect in the two Qwen2.5 settings; the null Thinking result and nonparallel historical sources still limit a source-independent interpretation.

We next ask whether external text absorbs TTE’s contribution. The same query and candidate entities are evaluated under era instructions, three target-era demonstrations, or retrieval-augmented generation (RAG) using the four highest-ranked

historical passages; each condition is compared with and without TTE. Table 10 reports the paired TLO changes. TTE is positive across all three comparisons for both Qwen2.5 models, while Qwen3-30B and Qwen3-4B-Think are mixed or null. PR does not improve overall, and context negative log-likelihood (NLL) increases. TTE therefore contributes an era-directed probability shift in the Qwen2.5 settings, while generation quality remains a separate objective.

A supervised adapter proceeds to held-out testing only after meeting a prespecified validation criterion. The Base and Thinking adapters meet this criterion; the Instruct adapter remains validation-only. On 180 held-out queries, TTE contributes an additional  $\Delta$ TLO of .425 [.322, .530] beyond the Base adapter, but  $-.292 [-.472, -.115]$  beyond the Thinking adapter. TTE is therefore complementary to the Base adapter and reverses under reasoning-time decoding, making adapter-relative effects architecture-dependent.

#### A.1.5 Blind Rubric and Norm Sensitivity

Two independent LLM judges, GPT-5.5 and Claude Fable 5, evaluate 120 pairs presented without condition labels and with randomized left–right order. The rubric covers target-era fit, topic preservation, future leakage, fluency, and visible-marker

| Model          | Controlled – prompt | Controlled – unpaired | $\Delta$ PR | $\Delta$ FLR |
|----------------|---------------------|-----------------------|-------------|--------------|
| Qwen-14B       | .445 [.373, .515]   | .021 [.014, .028]     | -.034       | -.105        |
| Qwen-14B-Inst  | .765 [.577, .964]   | -.048 [-.089, -.009]  | -.031       | -.056        |
| Qwen3-4B-Think | .132 [-.249, .513]  | .000 [-.017, .018]    | .048        | -.024        |

Table 9: Source-controlled corpus centroids ( $N = 180$  evaluation queries per model). The first two columns report paired  $\Delta$ TLO [95% CI] relative to prompting and to the unpaired corpus centroids; PR and FLR differences are relative to prompting.

| Model          | TTE – prompt         | TTE+3-shot – 3-shot | TTE+RAG – RAG       |
|----------------|----------------------|---------------------|---------------------|
| Qwen-14B       | .421 [.262, .580]    | .369 [.220, .517]   | .414 [.271, .557]   |
| Qwen-14B-Inst  | .765 [.213, 1.317]   | .507 [.002, 1.012]  | .448 [.047, .849]   |
| Qwen3-30B      | -.796 [-1.665, .072] | -.177 [-.410, .057] | .377 [-.237, .990]  |
| Qwen3-4B-Think | -.028 [-.334, .278]  | -.043 [-.383, .297] | -.032 [-.293, .228] |

Table 10: Paired  $\Delta$ TLO [95% CI] against text baselines ( $N = 180$  queries per model and comparison).

reliance. Claude records 43 TTE wins, 44 prompt wins, and 33 ties (exact  $p = 1$ ); agreement on the TTE/prompt/tie decision is .667 ( $\kappa = .490$ ). All 30 order-reversed repeats preserve the decision. The rubric therefore establishes cross-judge and presentation-order reliability, while aggregate preference remains balanced.

Because the intervention dose is proportional to the full L2 norm, one unusually large activation coordinate could inflate the applied change. For each model–language–era–method condition, p95 and p99 denote the 95th and 99th percentiles of its activation-norm distribution. Across 875 conditions, the ratio of the largest observed norm to that condition’s p99 has median, 95th-percentile, and maximum values of 1.009, 1.078, and 1.252. A separate 108-output test compares the unclipped rule with p95 and p99 caps: PR increases from .448 to .543 and .563, while FLR decreases from .070 to .061 and .048. Both caps preserve the direction of the effect in this sample; the primary results use the unclipped rule.

## A.2 Temporal Steering across Architectures and Methods

The main text uses Qwen2.5-14B-Base as a detailed example. This appendix asks how world-knowledge effects, style-residual behavior, and cross-lingual transfer change across model families, post-training regimes, and steering methods.

Tables 11–12 and 14–21 report mean  $\pm$  standard deviation across the three historical eras (*Old*, *Middle*, *Early Modern*). They compare the four

steering mechanisms defined in Section 3.3: CAA (Contrastive Activation Addition), Ens (Ensemble CAA), CMP (Chronological Manifold Projection), and EnsCMP (Ensemble Manifold Projection). Table 13 separately compares the full temporal vector with linear and nonlinear style residuals across model–language pairs.

### A.2.1 Epistemic Boundary Results

We first examine how steering geometry changes future leakage and era-appropriate retrieval. Tables 11 and 12 present the Future Leakage Rate (FLR) and Precision Rate (PR).

The results show a clear interaction between model geometry and steering method. For Qwen-14B, EnsCMP outperforms the discrete CAA baseline: it reduces FLR from  $0.338 \pm 0.357$  to  $0.260 \pm 0.319$  while improving PR from 0.362 to 0.607. Qwen3-30B achieves its best result on both metrics with CMP (FLR 0.061, PR 0.939), and Qwen3-4B-Think likewise attains its lowest FLR and highest PR under CMP (FLR 0.284, PR 0.700). Manifold projection therefore sharpens epistemic boundaries in several evaluated architectures.

The full model set also shows architecture-dependent estimator rankings. Qwen-14B-Inst performs best with the simpler ensemble method (FLR 0.299, PR 0.684), while Llama-8B-Inst and Gemma-9B-Inst remain strongest under the raw CAA baseline. Both manifold-aware and discrete methods therefore lead in different architectures and post-training regimes.

| Model          | CAA                                 | Ens                                 | CMP                                 | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| Qwen-14B       | $0.338 \pm 0.357$                   | $0.428 \pm 0.314$                   | $0.324 \pm 0.384$                   | <b><math>0.260 \pm 0.319</math></b> |
| Qwen-14B-Inst  | $0.370 \pm 0.279$                   | <b><math>0.299 \pm 0.211</math></b> | $0.364 \pm 0.246$                   | $0.356 \pm 0.291$                   |
| Llama-8B-Inst  | <b><math>0.187 \pm 0.237</math></b> | $0.307 \pm 0.289$                   | $0.316 \pm 0.435$                   | $0.420 \pm 0.386$                   |
| Gemma-9B-Inst  | <b><math>0.260 \pm 0.201</math></b> | $0.331 \pm 0.228$                   | $0.329 \pm 0.246$                   | $0.449 \pm 0.320$                   |
| Qwen3-30B      | $0.210 \pm 0.076$                   | $0.113 \pm 0.120$                   | <b><math>0.061 \pm 0.084</math></b> | $0.133 \pm 0.120$                   |
| Qwen3-4B-Think | $0.355 \pm 0.236$                   | $0.309 \pm 0.305$                   | <b><math>0.284 \pm 0.228</math></b> | $0.341 \pm 0.221$                   |

Table 11: Epistemic Cutoff results. FLR is the fraction of generated entities classified as later than the target era; lower is better.

| Model          | CAA                                 | Ens                                 | CMP                                 | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| Qwen-14B       | $0.362 \pm 0.367$                   | $0.505 \pm 0.352$                   | $0.543 \pm 0.411$                   | <b><math>0.607 \pm 0.354</math></b> |
| Qwen-14B-Inst  | $0.613 \pm 0.274$                   | <b><math>0.684 \pm 0.264</math></b> | $0.535 \pm 0.296$                   | $0.644 \pm 0.291$                   |
| Llama-8B-Inst  | <b><math>0.597 \pm 0.352</math></b> | $0.559 \pm 0.232$                   | $0.567 \pm 0.462$                   | $0.530 \pm 0.371$                   |
| Gemma-9B-Inst  | <b><math>0.740 \pm 0.201</math></b> | $0.652 \pm 0.235$                   | $0.621 \pm 0.217$                   | $0.518 \pm 0.286$                   |
| Qwen3-30B      | $0.790 \pm 0.076$                   | $0.888 \pm 0.120$                   | <b><math>0.939 \pm 0.084</math></b> | $0.717 \pm 0.231$                   |
| Qwen3-4B-Think | $0.595 \pm 0.245$                   | $0.691 \pm 0.305$                   | <b><math>0.700 \pm 0.280</math></b> | $0.642 \pm 0.227$                   |

Table 12: Epistemic Cutoff results. PR is the fraction of generated entities classified as era-appropriate; higher is better.

### A.2.2 Stability after Estimated Style Removal

To measure how much temporal control remains after removing an estimated style subspace, we compare the full time vector ( $v_{time}$ ) with the linear style residual ( $v_{cog}$ ).

Tables 14–15 and Tables 16–17 show that the estimated style subspace does not exhaust the temporal signal. Across the six models, the best-performing method under  $v_{time}$  typically remains competitive under  $v_{cog}$ , and in several cases the style residual is at least as effective as the full vector. For Qwen-14B, EnsCMP keeps FLR low when moving from  $v_{time}$  to  $v_{cog}$  ( $0.172 \rightarrow 0.179$ ) while increasing PR ( $0.737 \rightarrow 0.763$ ). Llama-8B-Inst shows the same stability: under EnsCMP, FLR decreases from 0.196 to 0.173 while PR rises from 0.770 to 0.810. These comparisons show that the measured epistemic boundary shift persists after removing the leading linear style components.

Architecture-specific differences remain visible. Qwen-14B-Inst is best under Ens both before and after style removal, while Gemma-9B-Inst consistently prefers CAA. In the Qwen family, Qwen3-30B reaches its highest PR under EnsCMP for both  $v_{time}$  and  $v_{cog}$  (0.891 and 0.886), and the residual gives the best FLR (0.114). Qwen3-4B-Think is also stable under EnsCMP, with FLR changing from 0.223 to 0.216 and PR from 0.777 to 0.776.

A residual temporal effect persists under this operational style removal, while estimator rankings remain architecture-dependent.

To test whether the result depends on the linear PCA assumption, we compare the linear PCA residual with the nonlinear MLP-autoencoder residual defined in Section 3.5. Table 13 reports mean PR over the three historical target eras for the full temporal vector, the PCA residual, and the AE residual. The AE residual improves over PCA in 6/12 model-language pairs, but the overall mean is slightly lower (0.281 vs. 0.289). Time-style coupling therefore persists under both linear and nonlinear style models.

### A.2.3 Cross-Lingual Temporal Transfer

We further examine cross-lingual temporal transfer by separating direct transfer, which applies source-language directions without alignment, from manifold-aligned transfer, which maps the source trajectory into target-language coordinates using the four era anchors.

The contrast between FLR (Tables 18, 19) and PR (Tables 20, 21) shows a structured pattern of cross-lingual temporal transfer. For Qwen-14B, EnsCMP is strongest under manifold-aligned transfer: relative to CAA, it reduces FLR from 0.701 to 0.196 while increasing PR from 0.146 to 0.371. Qwen-14B-Inst instead favors Ens in both settings

| Model          | Lang. | Full PR | PCA PR | AE PR | $\Delta_{\text{AE-PCA}}$ |
|----------------|-------|---------|--------|-------|--------------------------|
| Qwen-14B       | Zh    | 0.038   | 0.019  | 0.043 | +0.024                   |
| Qwen-14B       | En    | 0.383   | 0.384  | 0.359 | -0.025                   |
| Qwen-14B-Inst  | Zh    | 0.022   | 0.023  | 0.030 | +0.007                   |
| Qwen-14B-Inst  | En    | 0.571   | 0.548  | 0.543 | -0.005                   |
| Llama-8B-Inst  | Zh    | 0.257   | 0.239  | 0.245 | +0.006                   |
| Llama-8B-Inst  | En    | 0.214   | 0.228  | 0.196 | -0.032                   |
| Gemma-9B-Inst  | Zh    | 0.123   | 0.138  | 0.114 | -0.024                   |
| Gemma-9B-Inst  | En    | 0.661   | 0.672  | 0.669 | -0.003                   |
| Qwen3-30B      | Zh    | 0.010   | 0.013  | 0.016 | +0.003                   |
| Qwen3-30B      | En    | 0.615   | 0.598  | 0.526 | -0.072                   |
| Qwen3-4B-Think | Zh    | 0.060   | 0.061  | 0.077 | +0.016                   |
| Qwen3-4B-Think | En    | 0.555   | 0.546  | 0.551 | +0.006                   |
| Mean           | –     | 0.292   | 0.289  | 0.281 | -0.008                   |

Table 13: Nonlinear style-removal comparison across evaluated decoder models. Values are mean PR over three target eras for the full temporal vector, linear PCA residual, and nonlinear MLP-autoencoder residual. Positive  $\Delta_{\text{AE-PCA}}$  means the AE residual preserves more era-appropriate precision than the PCA residual.

| Model          | CAA                                 | Ens                                 | CMP               | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------|-------------------------------------|
| Qwen-14B       | $0.252 \pm 0.240$                   | $0.368 \pm 0.269$                   | $0.202 \pm 0.188$ | <b><math>0.172 \pm 0.232</math></b> |
| Qwen-14B-Inst  | $0.309 \pm 0.208$                   | <b><math>0.255 \pm 0.178</math></b> | $0.341 \pm 0.282$ | $0.323 \pm 0.238$                   |
| Llama-8B-Inst  | <b><math>0.140 \pm 0.153</math></b> | $0.285 \pm 0.200$                   | $0.244 \pm 0.216$ | $0.196 \pm 0.160$                   |
| Gemma-9B-Inst  | <b><math>0.162 \pm 0.154</math></b> | $0.188 \pm 0.175$                   | $0.202 \pm 0.117$ | $0.241 \pm 0.223$                   |
| Qwen3-30B      | <b><math>0.106 \pm 0.103</math></b> | $0.198 \pm 0.190$                   | $0.152 \pm 0.123$ | $0.109 \pm 0.107$                   |
| Qwen3-4B-Think | $0.371 \pm 0.135$                   | $0.339 \pm 0.165$                   | $0.269 \pm 0.132$ | <b><math>0.223 \pm 0.144</math></b> |

Table 14: Full-vector disentanglement baseline. FLR evaluates  $\mathbf{v}_{time}$  before estimated style removal; lower is better.

(direct: FLR 0.213, PR 0.527; manifold-aligned: FLR 0.240, PR 0.510). Temporal information therefore transfers in both models, although the preferred estimator changes after instruction tuning.

The remaining architectures show the same estimator dependence. Llama-8B-Inst obtains its lowest FLR with CAA in both direct and manifold-aligned transfer (0.111 and 0.150), but its highest PR with CMP directly (0.657) and EnsCMP after manifold-aligned transfer (0.512). Gemma-9B-Inst favors simpler estimators, Qwen3-30B favors Ens after manifold-aligned transfer, and Qwen3-4B-Think separates the method that minimizes direct FLR from the one that maximizes manifold-aligned PR. Across these architectures, estimator rankings vary by metric and alignment setting; no method dominates across model families.

### A.3 Held-Out Historical Distribution Correspondence

We use held-out perplexity (PPL) as an era-fit measure for broad-era adaptation. Time vectors are estimated on an 80% split of the Diachronic Steering Corpora and evaluated on the held-out 20% split; rows index Old, Middle, and Early Modern steering signals plus the unsteered baseline, and columns index held-out corpora from Old through Modern. Because raw PPL values are not directly comparable across tokenizers and model scales, each model-language matrix is column-normalized before averaging across the six evaluated decoder models.

Figure 7 reports the six-model mean using the same matrix layout as the single-model analysis. The baseline row remains strong in normalized PPL, reflecting the broad language-model fit of the unsteered decoder. Within the historical steering rows, however, same-era diagonal entries are

| Model          | CAA                                 | Ens                                 | CMP               | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------|-------------------------------------|
| Qwen-14B       | $0.294 \pm 0.295$                   | $0.386 \pm 0.273$                   | $0.184 \pm 0.249$ | <b><math>0.179 \pm 0.222</math></b> |
| Qwen-14B-Inst  | $0.317 \pm 0.248$                   | <b><math>0.265 \pm 0.185</math></b> | $0.351 \pm 0.308$ | $0.351 \pm 0.229$                   |
| Llama-8B-Inst  | <b><math>0.156 \pm 0.135</math></b> | $0.290 \pm 0.244$                   | $0.321 \pm 0.231$ | $0.173 \pm 0.162$                   |
| Gemma-9B-Inst  | <b><math>0.174 \pm 0.163</math></b> | $0.213 \pm 0.205$                   | $0.200 \pm 0.121$ | $0.263 \pm 0.217$                   |
| Qwen3-30B      | $0.129 \pm 0.115$                   | $0.179 \pm 0.217$                   | $0.179 \pm 0.144$ | <b><math>0.114 \pm 0.126</math></b> |
| Qwen3-4B-Think | $0.356 \pm 0.141$                   | $0.372 \pm 0.181$                   | $0.248 \pm 0.127$ | <b><math>0.216 \pm 0.151</math></b> |

Table 15: Linear style-residual results. FLR evaluates  $\mathbf{v}_{cog}$  after removing the estimated style subspace; lower is better.

| Model          | CAA                                 | Ens                                 | CMP               | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------|-------------------------------------|
| Qwen-14B       | $0.715 \pm 0.269$                   | $0.582 \pm 0.312$                   | $0.690 \pm 0.237$ | <b><math>0.737 \pm 0.298</math></b> |
| Qwen-14B-Inst  | $0.674 \pm 0.254$                   | <b><math>0.736 \pm 0.209</math></b> | $0.551 \pm 0.328$ | $0.610 \pm 0.290$                   |
| Llama-8B-Inst  | $0.760 \pm 0.224$                   | $0.682 \pm 0.172$                   | $0.748 \pm 0.217$ | <b><math>0.770 \pm 0.229</math></b> |
| Gemma-9B-Inst  | <b><math>0.838 \pm 0.154</math></b> | $0.803 \pm 0.192$                   | $0.798 \pm 0.117$ | $0.717 \pm 0.242$                   |
| Qwen3-30B      | $0.869 \pm 0.156$                   | $0.785 \pm 0.226$                   | $0.848 \pm 0.123$ | <b><math>0.891 \pm 0.107</math></b> |
| Qwen3-4B-Think | $0.629 \pm 0.135$                   | $0.661 \pm 0.165$                   | $0.723 \pm 0.169$ | <b><math>0.777 \pm 0.144</math></b> |

Table 16: Full-vector disentanglement baseline. PR evaluates era-appropriate entity retention under  $\mathbf{v}_{time}$ ; higher is better.

| Model          | CAA                                 | Ens                                 | CMP               | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------|-------------------------------------|
| Qwen-14B       | $0.672 \pm 0.329$                   | $0.598 \pm 0.280$                   | $0.658 \pm 0.314$ | <b><math>0.763 \pm 0.259</math></b> |
| Qwen-14B-Inst  | $0.625 \pm 0.318$                   | <b><math>0.736 \pm 0.185</math></b> | $0.574 \pm 0.337$ | $0.616 \pm 0.233$                   |
| Llama-8B-Inst  | $0.744 \pm 0.215$                   | $0.676 \pm 0.199$                   | $0.671 \pm 0.227$ | <b><math>0.810 \pm 0.194</math></b> |
| Gemma-9B-Inst  | <b><math>0.826 \pm 0.163</math></b> | $0.787 \pm 0.205$                   | $0.791 \pm 0.143$ | $0.729 \pm 0.211$                   |
| Qwen3-30B      | $0.846 \pm 0.172$                   | $0.812 \pm 0.221$                   | $0.821 \pm 0.144$ | <b><math>0.886 \pm 0.126</math></b> |
| Qwen3-4B-Think | $0.644 \pm 0.141$                   | $0.628 \pm 0.181$                   | $0.744 \pm 0.164$ | <b><math>0.776 \pm 0.176</math></b> |

Table 17: Linear style-residual results. PR evaluates era-appropriate entity retention under  $\mathbf{v}_{cog}$ ; higher is better.

lower than off-diagonal historical entries in both Chinese (0.476 vs. 0.652) and English (0.367 vs. 0.678). Thus the PPL result shows era-relative correspondence among historical steering signals while retaining the baseline as an explicit reference condition.

#### A.4 Competing-Hypothesis Tests across Architectures

We evaluate three accounts in raw activation space: an ordered temporal-distance account, a strict clustered-era account, and an ordered stepwise account. Table 22 reports results across the evaluated architectures. All model-language pairs have positive Mantel correlations, and 11/12 pairs favor the temporal-distance model over a strict cluster baseline. The stepwise comparison is more stringent: only 5/12 pairs favor temporal distance over an

ordered stepwise alternative. Broad-era order therefore recurs across architectures, while evidence for metric spacing beyond ordered steps is model- and language-dependent.

#### A.5 Held-Out Decade Projection across Architectures

To test whether chronological order extends beyond the four fitted anchors, we project unseen 1900s–1990s decade activations into the original geometry without refitting. Table 6 reports all model–language pairs and their multiplicity-adjusted results. Five pairs survive correction, three are positive but nonsignificant, and four are weak or negative. Aggregate decade order is positive but varies substantially across architectures and languages.

| Model          | CAA                                 | Ens                                 | CMP               | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------|-------------------------------------|
| Qwen-14B       | $0.450 \pm 0.360$                   | $0.459 \pm 0.389$                   | $0.219 \pm 0.266$ | <b><math>0.168 \pm 0.226</math></b> |
| Qwen-14B-Inst  | $0.379 \pm 0.359$                   | <b><math>0.213 \pm 0.259</math></b> | $0.280 \pm 0.286$ | $0.304 \pm 0.297$                   |
| Llama-8B-Inst  | <b><math>0.111 \pm 0.204</math></b> | $0.198 \pm 0.288$                   | $0.139 \pm 0.257$ | $0.124 \pm 0.186$                   |
| Gemma-9B-Inst  | $0.411 \pm 0.384$                   | <b><math>0.381 \pm 0.399</math></b> | $0.389 \pm 0.381$ | $0.454 \pm 0.432$                   |
| Qwen3-30B      | <b><math>0.142 \pm 0.246</math></b> | $0.148 \pm 0.229$                   | $0.178 \pm 0.281$ | $0.192 \pm 0.230$                   |
| Qwen3-4B-Think | <b><math>0.366 \pm 0.301</math></b> | $0.419 \pm 0.353$                   | $0.478 \pm 0.337$ | $0.450 \pm 0.331$                   |

Table 18: FLR under direct cross-lingual transfer of unaligned source-language directions.

| Model          | CAA                                 | Ens                                 | CMP               | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------|-------------------------------------|
| Qwen-14B       | $0.701 \pm 0.357$                   | $0.473 \pm 0.364$                   | $0.442 \pm 0.310$ | <b><math>0.196 \pm 0.213</math></b> |
| Qwen-14B-Inst  | $0.386 \pm 0.378$                   | <b><math>0.240 \pm 0.271</math></b> | $0.321 \pm 0.283$ | $0.312 \pm 0.261$                   |
| Llama-8B-Inst  | <b><math>0.150 \pm 0.280</math></b> | $0.157 \pm 0.288$                   | $0.341 \pm 0.405$ | $0.204 \pm 0.258$                   |
| Gemma-9B-Inst  | <b><math>0.474 \pm 0.401</math></b> | $0.610 \pm 0.445$                   | $0.536 \pm 0.433$ | $0.542 \pm 0.412$                   |
| Qwen3-30B      | $0.280 \pm 0.272$                   | <b><math>0.247 \pm 0.315</math></b> | $0.484 \pm 0.315$ | $0.340 \pm 0.266$                   |
| Qwen3-4B-Think | <b><math>0.383 \pm 0.340</math></b> | $0.401 \pm 0.337$                   | $0.471 \pm 0.354$ | $0.491 \pm 0.293$                   |

Table 19: FLR under manifold-aligned cross-lingual transfer.

| Model          | CAA                                 | Ens                                 | CMP                                 | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| Qwen-14B       | $0.310 \pm 0.338$                   | $0.354 \pm 0.353$                   | $0.174 \pm 0.248$                   | <b><math>0.275 \pm 0.271</math></b> |
| Qwen-14B-Inst  | $0.271 \pm 0.365$                   | <b><math>0.527 \pm 0.359</math></b> | $0.327 \pm 0.381$                   | $0.386 \pm 0.295$                   |
| Llama-8B-Inst  | $0.492 \pm 0.413$                   | $0.468 \pm 0.405$                   | <b><math>0.657 \pm 0.403</math></b> | $0.583 \pm 0.382$                   |
| Gemma-9B-Inst  | <b><math>0.396 \pm 0.381</math></b> | $0.349 \pm 0.380$                   | $0.286 \pm 0.365$                   | $0.227 \pm 0.331$                   |
| Qwen3-30B      | $0.541 \pm 0.332$                   | <b><math>0.682 \pm 0.360</math></b> | $0.525 \pm 0.393$                   | $0.658 \pm 0.336$                   |
| Qwen3-4B-Think | <b><math>0.486 \pm 0.378</math></b> | $0.481 \pm 0.357$                   | $0.356 \pm 0.387$                   | $0.387 \pm 0.379$                   |

Table 20: PR under direct cross-lingual transfer of unaligned source-language directions.

| Model          | CAA                                 | Ens                                 | CMP               | EnsCMP                              |
|----------------|-------------------------------------|-------------------------------------|-------------------|-------------------------------------|
| Qwen-14B       | $0.146 \pm 0.228$                   | $0.294 \pm 0.325$                   | $0.189 \pm 0.237$ | <b><math>0.371 \pm 0.251</math></b> |
| Qwen-14B-Inst  | $0.348 \pm 0.387$                   | <b><math>0.510 \pm 0.345</math></b> | $0.356 \pm 0.363$ | $0.431 \pm 0.329$                   |
| Llama-8B-Inst  | $0.320 \pm 0.400$                   | $0.446 \pm 0.411$                   | $0.366 \pm 0.370$ | <b><math>0.512 \pm 0.407</math></b> |
| Gemma-9B-Inst  | <b><math>0.289 \pm 0.336</math></b> | $0.130 \pm 0.270$                   | $0.214 \pm 0.330$ | $0.178 \pm 0.304$                   |
| Qwen3-30B      | $0.494 \pm 0.353$                   | <b><math>0.527 \pm 0.419</math></b> | $0.359 \pm 0.328$ | $0.497 \pm 0.349$                   |
| Qwen3-4B-Think | $0.476 \pm 0.367$                   | <b><math>0.486 \pm 0.357</math></b> | $0.349 \pm 0.379$ | $0.405 \pm 0.343$                   |

Table 21: PR under manifold-aligned cross-lingual transfer.

## A.6 Unseen Intermediate Coordinates across Architectures

We next evaluate interpolated coordinates that were not used as training anchors. For each model and language, we compute the classifier probability assigned to the four anchor eras after steering with the manifold vector at  $t = 0.5$  and  $t = 1.5$ . Table 23 reports results across the evaluated architectures. The mixed distributions are consis-

tent with intermediate coordinates expressing mixtures of adjacent eras. The direction of movement varies by model and language, indicating that intermediate-coordinate behavior is probabilistic and architecture-dependent.

## A.7 Sparse Autoencoder Baseline

A sparse autoencoder (SAE) baseline asks whether a small set of features can replace TTE’s distributed

| Model          | Lang. | Mantel $r$ | $\Delta\text{BIC}$ strict | $\Delta\text{BIC}$ stepwise |
|----------------|-------|------------|---------------------------|-----------------------------|
| Qwen-14B       | Zh    | 0.341      | 172.0                     | 14.9                        |
| Qwen-14B       | En    | 0.602      | 564.2                     | -4.3                        |
| Qwen-14B-Inst  | Zh    | 0.448      | 241.5                     | -27.1                       |
| Qwen-14B-Inst  | En    | 0.908      | 2019.4                    | -989.9                      |
| Llama-8B-Inst  | Zh    | 0.538      | 564.2                     | 12.1                        |
| Llama-8B-Inst  | En    | 0.798      | -1012.5                   | -1807.3                     |
| Gemma-9B-Inst  | Zh    | 0.569      | 538.4                     | 9.2                         |
| Gemma-9B-Inst  | En    | 0.763      | 857.7                     | -188.8                      |
| Qwen3-30B      | Zh    | 0.699      | 982.0                     | -27.2                       |
| Qwen3-30B      | En    | 0.892      | 1675.0                    | -844.7                      |
| Qwen3-4B-Think | Zh    | 0.751      | 1466.2                    | 15.9                        |
| Qwen3-4B-Think | En    | 0.835      | 1109.7                    | 3.5                         |

Table 22: Competing-hypothesis tests across evaluated architectures. Positive  $\Delta\text{BIC}$  means the temporal-distance model fits better than the corresponding alternative.

| Model          | Lang. | $t = 0.5$ |        |         |        | $t = 1.5$ |        |         |        |
|----------------|-------|-----------|--------|---------|--------|-----------|--------|---------|--------|
|                |       | Old       | Middle | E. Mod. | Modern | Old       | Middle | E. Mod. | Modern |
| Qwen-14B       | Zh    | 0.313     | 0.265  | 0.244   | 0.178  | 0.290     | 0.257  | 0.238   | 0.216  |
| Qwen-14B       | En    | 0.341     | 0.239  | 0.174   | 0.246  | 0.223     | 0.343  | 0.209   | 0.225  |
| Qwen-14B-Inst  | Zh    | 0.317     | 0.251  | 0.244   | 0.188  | 0.311     | 0.251  | 0.250   | 0.188  |
| Qwen-14B-Inst  | En    | 0.180     | 0.178  | 0.252   | 0.391  | 0.198     | 0.226  | 0.227   | 0.348  |
| Llama-8B-Inst  | Zh    | 0.315     | 0.244  | 0.233   | 0.207  | 0.301     | 0.251  | 0.235   | 0.213  |
| Llama-8B-Inst  | En    | 0.293     | 0.172  | 0.213   | 0.322  | 0.144     | 0.192  | 0.277   | 0.388  |
| Gemma-9B-Inst  | Zh    | 0.243     | 0.251  | 0.274   | 0.231  | 0.252     | 0.255  | 0.268   | 0.226  |
| Gemma-9B-Inst  | En    | 0.383     | 0.187  | 0.168   | 0.262  | 0.208     | 0.215  | 0.241   | 0.335  |
| Qwen3-30B      | Zh    | 0.307     | 0.280  | 0.231   | 0.183  | 0.300     | 0.263  | 0.241   | 0.195  |
| Qwen3-30B      | En    | 0.144     | 0.169  | 0.288   | 0.399  | 0.147     | 0.209  | 0.290   | 0.354  |
| Qwen3-4B-Think | Zh    | 0.275     | 0.282  | 0.242   | 0.202  | 0.214     | 0.264  | 0.276   | 0.247  |
| Qwen3-4B-Think | En    | 0.180     | 0.189  | 0.255   | 0.376  | 0.172     | 0.219  | 0.257   | 0.352  |

Table 23: Soft era probabilities for unseen intermediate coordinates across evaluated architectures.

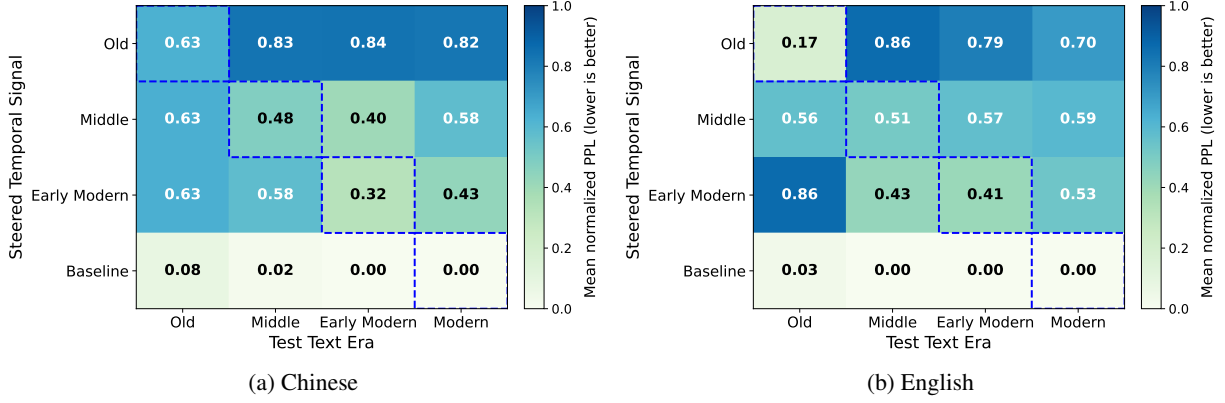


Figure 7: Six-model mean held-out PPL matrices. Rows index the steering era, including the unsteered baseline, and columns index the held-out test-corpus era. Each model-language matrix is column-normalized before averaging; lower values indicate better fit.

| Model          | WK-FLR | WK-PR | XL-PR |
|----------------|--------|-------|-------|
| Qwen-14B       | 0.633  | 0.367 | 0.244 |
| Qwen-14B-Inst  | 0.665  | 0.335 | 0.179 |
| Llama-8B-Inst  | 0.717  | 0.283 | 0.100 |
| Gemma-9B-Inst  | 0.739  | 0.261 | 0.442 |
| Qwen3-30B      | 0.632  | 0.000 | 0.000 |
| Qwen3-4B-Think | 0.805  | 0.000 | 0.000 |

Table 24: SAE baseline aggregate metrics. WK-FLR/WK-PR evaluate world-knowledge probing; XL-PR evaluates cross-lingual precision.

| Model          | Cond. | Layers | FLR   | PR    |
|----------------|-------|--------|-------|-------|
| Qwen-14B       | R1    | 16     | 0.000 | 1.000 |
| Qwen-14B       | R2    | 16     | 0.067 | 0.933 |
| Qwen3-4B-Think | R1    | 14     | 0.222 | 0.736 |
| Qwen3-4B-Think | R2    | 14     | 0.158 | 0.801 |
| Gemma-9B-Inst  | R1    | 14     | 0.078 | 0.922 |
| Gemma-9B-Inst  | R2    | 14     | 0.275 | 0.725 |

Table 25: Temporal metrics after reinitializing selected layers (R1) or nonselected layers at comparable depths (R2).

residual directions. We train a 4096-feature over-complete SAE for 50 epochs with an L1 sparsity penalty, select the 50 features most informative about era on the same layers used for steering, and evaluate the resulting direction on the Epistemic Cutoff and cross-lingual tasks. In Table 24, WK-FLR and WK-PR denote future leakage and era-appropriate precision on the world-knowledge task, while XL-PR denotes precision under cross-lingual transfer.

Sparse-feature precision is unstable across models: Qwen3-30B and Qwen3-4B-Think have zero WK-PR and XL-PR, Llama-8B-Inst has low XL-PR, and Qwen-14B-Inst remains below the base model in cross-lingual precision. Gemma-9B-Inst is the strongest sparse-feature case, but its world-knowledge precision remains modest. Sparse features therefore capture parts of the temporal signal without reliably replacing the distributed directions used by TTE.

## A.8 Randomized Model Controls

Randomized-model controls ask whether the evaluation can assign favorable temporal scores after the relevant pretrained structure has been destroyed. R1 reinitializes the union of layers selected as temporally informative, whereas R2 reinitializes nonselected layers at comparable depths. We then repeat the reduced manifold analysis.

Table 25 shows that FLR/PR can remain superficially favorable after either randomization. The corresponding generations, however, contain repeated multilingual fragments, markup-like strings, and degenerate lexical loops; none is coherent and era-appropriate. Coherent steering therefore depends on the pretrained layer structure targeted by TTE, and FLR/PR cannot be interpreted without output quality in this control.

## A.9 Encoder and Recurrent Architecture Controls

Encoder-only and recurrent controls ask whether exposure to temporally labeled corpora is suffi-

cient to reproduce the observed behavior. We train architecture-appropriate probes for mBERT and a four-layer BiLSTM masked language model on the same corpora. Because neither model supports comparable free-form generation, FLR and PR are measured through masked-token analogues; alignment is measured by centered kernel alignment (CKA).

Both controls separate era labels, but neither recovers the corresponding temporal behavior (Table 26). mBERT reaches perfect probe accuracy but only 0.017 PR, while the BiLSTM has high probe accuracy and near-zero PR across steering variants. Era information can therefore be readable without reproducing the generative temporal control observed in the decoder models.

| Control       | Probe | FLR   | PR    | Align. |
|---------------|-------|-------|-------|--------|
| mBERT         | 1.000 | 0.002 | 0.017 | 0.659  |
| BiLSTM CAA    | 0.963 | 0.003 | 0.000 | 0.409  |
| BiLSTM Ens    | 0.966 | 0.006 | 0.000 | 0.409  |
| BiLSTM CMP    | 0.964 | 0.003 | 0.000 | 0.409  |
| BiLSTM EnsCMP | 0.966 | 0.008 | 0.001 | 0.409  |

Table 26: Architecture controls. Alignment is best CKA for mBERT and best cross-lingual CKA for BiLSTM.

### A.10 Prompt-Conflict Stress Test

Prompt-conflict tests ask whether temporal steering can still shape generation when the user instruction explicitly requests a contemporary register. We evaluate parallel English and Chinese queries that request modern-teenager, internet-slang, or texting-style answers while applying Old, Middle, or Early Modern steering vectors. Representative generations are shown in Tables 27 and 28. Across the six evaluated models, usable era-conditioned override appears for all three target eras in both languages. Old and Middle steering are more stable than Early Modern steering, and Qwen3-4B-Thinking shows weaker overrides under reasoning-time decoding. These results indicate that residual-stream temporal control can compete with explicit stylistic instructions, with strength shaped by target era, architecture, and decoding regime.

### A.11 Cross-Lingual Temporal Transfer and Language Coupling

Table 5 in the main text and Table 29 below hold the query fixed while varying the language and era of the temporal direction. Native transfer uses

a direction estimated in the query language (e.g., ZH→ZH), whereas cross-lingual transfer uses a direction from the other language (e.g., EN→ZH). The quantitative analysis further distinguishes direct transfer from manifold-aligned transfer, which maps the source trajectory into target-language coordinates before steering.

In Table 29, an Old English direction applied to a Modern Chinese query elicits Old English, and later English directions elicit the corresponding English historical registers. Together with the converse English-query example in Table 5, this output-language override illustrates coupling between temporal and linguistic features. The FLR/PR experiments in Section A.2.3 quantify the effect and show that its strongest estimator varies across architectures.

### A.12 Model details

Table 30 reports the architectural dimensions needed to reproduce the six-model analyses.

### A.13 Artifacts, Licenses, and Compute

The scientific artifacts used in this work are the pretrained model checkpoints in Table 30, the diachronic steering corpora described in Section 3.2, the era-character templates and evaluation datasets documented in Appendices A.14 and A.16, and the generated outputs and analysis scripts released with the public repository. We cite the model creators in Table 30. Artifact license and redistribution terms for external model checkpoints, corpora, generated outputs, derived statistics, and third-party packages are documented at <https://github.com/LilTiger/Time-Travel-Engine>. The artifacts produced by this work are intended for research reproduction and analysis of temporal representations, not for deployment as historical fact sources.

All experiments use existing model weights; we do not pretrain, instruction-tune, or release new LLM checkpoints. Experiments were run on a single Apple-Silicon workstation using PyTorch MPS acceleration with 512GB unified memory. The complete study required approximately 344 single-device accelerator-hours, including six-model generation, robustness and sensitivity controls, text and adapter baselines, and sparse-feature analyses; lightweight probes and autoencoders account for a smaller share. Auxiliary analyses use the hyperparameters stated in the method sections: top-five steering layers, a 4096-feature SAE trained for 50 epochs, and a bottleneck- $k=5$  MLP autoen-

| Conflict prompt                                                                | Old steering                                                                                                                                                                                                                             | Middle steering                                                                                                                                                                                                        | Early Modern steering                                                                                                                                                                                                        |
|--------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Speak like a modern teenager, what is justice?                                 | Rihtwisnes is þæt aelc mann hæbbe riht dom, and ne bið nan mann forðon þurh meaht, feoh, oððe freondscipe.<br><i>Justice means that each person receives rightful judgment and is not ruined by power, wealth, or friendship.</i>        | Justice is that ech man have his right, and no man be wronged by power, favour, or fals judgement.<br><i>Justice means that each person has their right, and no one is wronged by power, favor, or false judgment.</i> | Iustice is the due order of right, whereby every person is iudged without favour, and wrong is repaired.<br><i>Justice is the due order of right, by which every person is judged without favor and wrongs are repaired.</i> |
| Answer in modern internet slang, but still answer directly: How was time kept? | Menn gemetton tid mid sunnan sceadwe, waeter-clocce, candeale, and gemearcodum tidum, swa daeg and niht hweorfodon.<br><i>People measured time by the sun's shadow, water clocks, candles, and marked hours as day and night turned.</i> | In olde tyme, men kepten the houres by sonne-dial, water-clock, candel, and hour-glasse.<br><i>In old times, people kept the hours by sundials, water clocks, candles, and hourglasses.</i>                            | Time was kept by sun-dials, hour-glasses, church-bells, and mechanick clocks as they came into use.<br><i>Time was kept by sundials, hourglasses, church bells, and mechanical clocks as they came into use.</i>             |
| Reply like a modern teenager texting a friend: How was food preserved?         | Mete and corn waeron gehealden mid sealte, smoc, dryunge, piclunge, and cealdum stowe, þæt hit ne rotode.<br><i>Meat and grain were kept with salt, smoke, drying, pickling, and cold storage so they would not rot.</i>                 | Folk kepte mete by saltinge, drying, smoking, and pickling it in vinegre or brine.<br><i>People preserved food by salting, drying, smoking, and pickling it in vinegar or brine.</i>                                   | Victuals were preserued by salting, smoking, drying, pickling, sugaring, and keeping them in coole cellars.<br><i>Food was preserved by salting, smoking, drying, pickling, sugaring, and storing it in cool cellars.</i>    |

Table 27: Prompt-conflict generations under historical steering; gray italics give translations.

coder trained for 100 epochs. The implementation uses PyTorch, Hugging Face Transformers, scikit-learn, SciPy, and NumPy; package-level details and command-line settings are provided in the released scripts.

The data consist of published or publicly accessible historical texts, curated prompts, and model generations. We did not collect private user data, recruit participants, or use crowdworkers. During curation, we used only textual content needed for temporal analysis and excluded person-level metadata not required for reproduction. Because historical sources may contain archaic or biased language, examples serve as evidence about model representations; cultural or period-level normative interpretation lies outside their use.

#### A.14 Queries for Temporal State Extraction

CAA varies the era condition while holding task content constant. Tables 31 and 32 list the parallel Chinese and English queries used for this contrast; the corresponding *Era Charters* specify the historical condition. The complete charters are available in our repository at <https://github.com/LilTiger/Time-Travel-Engine>.

#### A.15 Contrastive Style Dataset

To examine the relation between linguistic register and temporal boundary signals, we curated a contrastive style dataset. Its parallel pairs are designed to preserve core meaning while shifting register across four broad eras: *Old*, *Middle*, *Early Modern*, and *Modern*.

These pairs define the estimated style subspace ( $U_{style}$ ) used in the disentanglement experiments (Section 3.5). We apply PCA to the difference vectors between modern and archaic representations and retain the leading register-related directions.

Table 33 presents representative examples from the dataset, spanning diverse topics such as time, nature, and social customs. The full dataset, containing paired prose and verse for both Chinese and English, is available at <https://github.com/LilTiger/Time-Travel-Engine>.

#### A.16 Details of the Evaluation Benchmarks

This appendix describes how the three evaluation benchmarks instantiate complementary questions about temporal control and provides representative evaluation queries. The Epistemic Cut-off Dataset measures knowledge boundaries, the Causal Remodeling Dataset tests era-compatible explanatory resources, and the Mismatch Entan-

| Conflict prompt                                                                                                           | Old steering                                                                                                                                                                                                                  | Middle steering                                                                                                                                                                                                          | Early Modern steering                                                                                                                                                                                                                     |
|---------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 请用现代年轻人的口吻，像发朋友圈一样回答：什么是公道？<br><i>Use a contemporary youth voice, as if posting to social media: what is justice?</i>     | 公道者，乃人之常行也。何谓常行？常以正直为心，不偏不倚是也。如日月照临，昭昭然明；如江河奔流，澹澹兮平。<br><i>Justice is the common way people should follow: to keep an upright heart and judge without bias.</i>                                                               | 公道者，非独明察之士所识也。凡人皆有其心之所向，而此心之所向即为公道。然则何以知其所向乎？曰：“行己有耻。”<br><i>Justice is not known only by the perceptive. Each person has an inward sense of what is right, and one knows it by acting with shame before wrongdoing.</i> | 公道者，非独官府所秉之理也。凡人行事处事，皆应以公平公正之心为之，则谓之为公道矣。然则如何方能行此公道乎？当先察其情状，明辨是非曲直。<br><i>Justice is not only a principle held by officials. In human affairs, one should act with fairness, examine the circumstances, and discern right from wrong.</i> |
| 请用网络流行语直接回答：古人通常怎么记录时间？<br><i>Answer directly in internet slang: how did ancient people usually keep time?</i>            | 古之人，盖亦从其时令，以正岁月。<br><i>Ancient people followed seasonal ordinances to regulate months and years.</i>                                                                                                                          | 古人常用漏壶、铜钲等计时器，以及日晷、星宿等天文观测法来记时。<br><i>People commonly used clepsydras, bronze timing devices, sundials, and observations of the stars to keep time.</i>                                                                  | 此乃古之遗风，多赖乎漏水时辰，焚香度日，或晷仪测天道也。简言之，水钟、漏刻、香篆、日晷是常用。<br><i>This was an old practice: time was kept by flowing water, burning incense, or sundial instruments; water clocks, clepsydras, incense seals, and sundials were common.</i>           |
| 请像现代年轻人发微信一样回答：食物通常怎样保存？<br><i>Reply like a contemporary youth sending a chat message: how is food usually preserved?</i> | 存食之道，有三焉：一曰低温藏之，冰室是用；二曰晒干曝之，风日之下晒焉；三曰腌渍渍之，盐酒醋汁以调。<br><i>There are three ways to store food: keep it cool in an ice room, dry it under wind and sun, or preserve it by pickling with salt, vinegar, wine, or other brines.</i> | 盐腌：以鱼为本，惟旨故美。糖渍：梅实之味甘酸宜人，而未熟者苦涩难食；今则摘其青色初成者，投于缸中，杂以白糖。<br><i>Salting is used for fish; sugaring suits plums, whose sweet and sour flavor becomes pleasant when the young fruit is placed in a jar with sugar.</i>        | 有四事须具之，安得不教。一陈皮燥，以水浸之；二干肉，先用盐卤之，然后晒干；三米糠，先淘洗净。<br><i>Several methods are needed: dried tangerine peel is soaked, dried meat is first salted and then sun-dried, and rice bran is washed before use.</i>                                   |

Table 28: Chinese prompt-conflict generations under historical steering; gray italics give translations.

gument Dataset places contemporary requests in conflict with historical steering. The complete benchmark data and generation scripts are available in the public repository at <https://github.com/LilTiger/Time-Travel-Engine>.

### A.16.1 Epistemic Cutoff Dataset

This dataset probes chronological knowledge boundaries through open-ended queries requesting lists of entities (e.g., philosophers, inventions, tools), with representative examples in Table 34. The reference entity lists are segmented by broad historical era. For instance, when the model is steered to the *Old* era (e.g., Pre-Qin), an entity assigned to the *Modern* era (e.g., "iPhone" or "Kant") counts as future leakage, while entities assigned to or before the target era count toward PR.

### A.16.2 Causal Remodeling Dataset

The Causal Remodeling Dataset evaluates whether generated explanations use technologies and causal resources compatible with the target era. In contrast to entity-list retrieval, it presents practical scenarios involving lighting, communication, and preservation; sample queries appear in Table 35. FLR measures the appearance of later technologies (e.g., a “flashlight” under Old-era steering), while PR measures the use of benchmark-listed era-compatible mechanisms (e.g., “oil lamps”).

### A.16.3 Mismatch Entanglement Dataset

This dataset tests style–epistemic entanglement when evaluation queries request modern content or style. The queries pair modern cues with *Old*, *Middle*, or *Early Modern* steering; representative examples are shown in Table 36. We treat eras as broad temporal conditions and evaluate whether generations replace later-only resources with era-compatible concepts, technologies, and rhetorical conventions while preserving the requested task content. The examples vary the type of conflict: medical treatment and storage media create strong pressure toward contemporary knowledge, whereas sports, water systems, and painting materials probe everyday institutions and artifacts. This separates surface compliance with modern wording from changes in the era-compatible resources expressed in the output.

| PROMPT                                                                                                                                                                                              |                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                   |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 请以诗歌的形式描述一场冬日的暴风雪。<br><i>Please describe a winter storm in the form of a poem.</i>                                                                                                                  |                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                   |
| Baseline (Modern State)                                                                                                                                                                             | Native Direction (ZH→ZH)                                                                                                                                                                                                                                           | Cross-Lingual Direction (EN→ZH)                                                                                                                                                                                                                                                   |
| <b>Direction: Old</b>                                                                                                                                                                               |                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                   |
| 冬夜寒风起，雪花漫天飞。<br>银装素裹中，世界静悄悄...<br><i>Winter winds rise in the night,<br/>           snowflakes fly across the sky.<br/>           Amidst the silver-clad world, all<br/>           is silent...</i> | 朔风怒号兮，天地昏黄，素羽纷飞兮，六出其芒...<br><i>The north wind howls; heaven and earth grow dim. White feathers whirl, each bearing six rays...</i>                                                                                                                                 | þurfe geong wundorlic þing,<br>wolcenum geweorpeð on wolde,<br>hrerende hræfena gehyngum<br>hrimbealo sceawade.<br><i>A new wondrous thing may be needed; it comes to pass in the clouds over the plain, stirring the ravens' hunger as frost-ruin appears.</i>                   |
| <b>Direction: Middle</b>                                                                                                                                                                            |                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                   |
| 冬夜寒风起，雪花漫天飞。<br>银装素裹中，世界静悄悄...<br><i>Winter winds rise in the night,<br/>           snowflakes fly across the sky.<br/>           Amidst the silver-clad world, all<br/>           is silent...</i> | 朔风卷地雪漫天，万木凋零玉絮旋。冻合江河冰作岸，云迷岩谷霰连绵...<br><i>The north wind sweeps the earth as snow fills the sky; ten thousand trees wither while jade-like flakes swirl. Ice seals the rivers into frozen banks; clouds obscure the rocky valleys as sleet falls without end...</i> | The wynter storme is so wode,<br>The snow floures ful thikke<br>falleth, As hevy stones down<br>they shalle...<br><i>The winter storm is so furious;<br/>           The snow flowers fall full thick,<br/>           As heavy stones down they shall<br/>           [fall]...</i> |
| <b>Direction: Early Modern</b>                                                                                                                                                                      |                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                   |
| 冬夜寒风起，雪花漫天飞。<br>银装素裹中，世界静悄悄...<br><i>Winter winds rise in the night,<br/>           snowflakes fly across the sky.<br/>           Amidst the silver-clad world, all<br/>           is silent...</i> | 纷纷扬扬漫天，素裹银装玉栏，风卷残云舞翩跹...<br><i>Flurries fill the sky; jade railings are wrapped in silver-white. The wind sweeps lingering clouds into a whirling dance...</i>                                                                                                     | Thou art beholding unto<br>Winter's fierce tempest, As<br>þnow doth þwiftly þwell from<br>heaven above...<br><i>Behold Winter's fierce tempest,<br/>           as snow swiftly swells from<br/>           heaven above...</i>                                                     |

Table 29: Qualitative cross-lingual register modulation on Qwen2.5-14B-Base using a Chinese query. The baseline is compared with native-language and cross-lingual temporal directions; gray italics give English translations.

| Model                                            | Params          | Layers | Attention Heads | Hidden States | Feedforward       |
|--------------------------------------------------|-----------------|--------|-----------------|---------------|-------------------|
| Qwen2.5-14B-Base (Qwen Team et al., 2024)        | 14B             | 48     | 40              | 5120          | 13824             |
| Qwen2.5-14B-Instruct (Qwen Team et al., 2024)    | 14B             | 48     | 40              | 5120          | 13824             |
| Llama-3.1-8B-Instruct (Grattafiori et al., 2024) | 8B              | 32     | 32              | 4096          | 14336             |
| Gemma-2-9B-Instruct (Gemma Team et al., 2024)    | 9B              | 42     | 16              | 3584          | 14336             |
| Qwen3-30B-A3B-Instruct (Yang et al., 2025)       | 30B / 3B active | 48     | 32              | 2048          | 6144 / 768 expert |
| Qwen3-4B-Thinking (Yang et al., 2025)            | 4B              | 36     | 32              | 2560          | 9728              |

Table 30: Model specifications. For Qwen3-30B-A3B-Instruct, the feedforward column reports shared and per-expert intermediate dimensions.

| #  | Query                        |
|----|------------------------------|
| 1  | 用三句话介绍你自己、你的身份和你每日所见的制度。     |
| 2  | 描写春日清晨城郭或乡野的景色。              |
| 3  | 解释什么是‘薄如纸而能映人像与远景，可收于衣囊’的怪物。 |
| 4  | 说明如何度量时间、距离、重量，常用何种单位？       |
| 5  | 城中多人发热咳嗽而亡，你如何判断病因并建议官府采取措施？ |
| 6  | 连续歉收或大水成灾，你认为成因为何？如何赈济安民？    |
| 7  | 描述一次远行的准备工作和沿途可能遇到的困难。       |
| 8  | 解释天象变化（日食、彗星）的含义和应对方法。       |
| 9  | 介绍当世最重要的三部典籍及其核心思想。          |
| 10 | 描述一场婚礼或丧葬仪式的基本流程。            |
| 11 | 说明如何治疗常见的风寒发热之症。             |
| 12 | 解释为何某些行为会被视为违礼或犯法。           |
| 13 | 描述一位德高望重之人应具备的品质。            |
| 14 | 说明农事的基本节气安排和耕作要点。            |
| 15 | 描写月夜独坐或登高远眺的感受（诗意表达）。        |

Table 31: Chinese queries held constant across the four era conditions.

| #  | Query                                                                                                    |
|----|----------------------------------------------------------------------------------------------------------|
| 1  | Introduce yourself in three sentences: your identity and the institutions you encounter daily.           |
| 2  | Describe a spring morning scene near a town or countryside.                                              |
| 3  | Explain what a 'paper-thin object that reflects faces and distant scenes, fitting in a pocket' might be. |
| 4  | Explain how you measure time, distance, and weight. What units are commonly used?                        |
| 5  | A fever and coughing illness kills many in town. How do you determine the cause and advise authorities?  |
| 6  | Continuous crop failure or flooding occurs. What causes do you suspect and how to provide relief?        |
| 7  | Describe preparations for a long journey and difficulties you might encounter.                           |
| 8  | Explain the meaning of celestial changes (eclipses, comets) and how to respond.                          |
| 9  | Introduce the three most important texts of your era and their core ideas.                               |
| 10 | Describe the basic procedures of a wedding or funeral ceremony.                                          |
| 11 | Explain how to treat a common cold or fever.                                                             |
| 12 | Explain why certain behaviors are considered improper or illegal.                                        |
| 13 | Describe the qualities a person of great virtue should possess.                                          |
| 14 | Explain the basic agricultural calendar and farming essentials.                                          |
| 15 | Describe feelings while sitting alone on a moonlit night or gazing from a height (poetic expression).    |

Table 32: English queries held constant across the four era conditions.

| Period                                   | Chinese Text                     | English Text                                                                                                                                  |
|------------------------------------------|----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Topic: Passage of Time</i>            |                                  |                                                                                                                                               |
| <i>Old</i>                               | 逝者如斯夫，不舍昼夜。                      | Tid ne bideþ nænigne mann, hēo æfre forðgæþ.                                                                                                  |
| <i>Middle</i>                            | 光阴似箭，日月如梭，岁不我与，时不再来。             | Tyme abideth noon, it passeth as a stream.                                                                                                    |
| <i>E. Mod</i>                            | 俗话说得好，一寸光阴一寸金，寸金难买寸光阴哪。          | Time and Tide wait for no Man, but ever flow onward.                                                                                          |
| <i>Modern</i>                            | 时间不等人。                           | Time waits for no one.                                                                                                                        |
| <i>Topic: Blooming Flowers</i>           |                                  |                                                                                                                                               |
| <i>Old</i>                               | 百花竞放，其香四溢。                       | Blostman springaþ, hira wlite is fæger.                                                                                                       |
| <i>Middle</i>                            | 繁花似锦，争奇斗艳，满园春色关不住。               | The floures bloomen faire, of dyverse colours and swete smell.                                                                                |
| <i>E. Mod</i>                            | 花儿都开了，红的白的紫的，好不热闹，香气扑鼻。          | The Flowers are blooming in great abundance, their sweet perfume fills the air.                                                               |
| <i>Modern</i>                            | 花开了。                             | The flowers are blooming.                                                                                                                     |
| <i>Topic: Autumn Scenery (Verse)</i>     |                                  |                                                                                                                                               |
| <i>Old</i>                               | 蒹葭苍苍，白露为霜。秋风萧瑟，草木摇落。             | Hærfest is cumen, hēah wind blæwþ; lēaf feallaþ, lond wearþ cald.                                                                             |
| <i>Middle</i>                            | 碧云天外雁南飞，黄叶纷纷满地堆。萧瑟西风吹不尽，一年秋色到楼台。 | The leves fallen from the tre, The wynde doth blowe so colde and kene; The somer dayes now fledde be, And al the worlde is bare and lene.     |
| <i>E. Mod</i>                            | 西风吹落叶，黄叶满庭阶。一年好景处，秋色最堪怀。         | When Autumn comes with golden Hue, The Leaves do fall from every Tree; The Summer Days have bid Adieu, And Nature mourns what used to be.     |
| <i>Modern</i>                            | 秋叶飘落，铺满了小径。风吹过树梢，带走了夏天。          | Autumn leaves are falling, carpeting the ground in gold. The wind whispers through bare branches, a farewell to warmth.                       |
| <i>Topic: Wine and Gathering (Verse)</i> |                                  |                                                                                                                                               |
| <i>Old</i>                               | 我有嘉宾，鼓瑟吹笙。呦呦鹿鸣，食野之苹。             | Wīn bringþ glædnesse, gāstas āhebbað; drync and drēam, dēorwyrðe sæl.                                                                         |
| <i>Middle</i>                            | 葡萄美酒夜光杯，欲饮琵琶马上催。醉卧沙场君莫笑，古来征战几人回。 | Wyn bryngeth joye and maketh glad, It lyghteth up the derke of nyght; It comforteth the herte that's sad, And turneth wronge to alle aryght.  |
| <i>E. Mod</i>                            | 劝君更尽一杯酒，西出阳关无故人。今朝有酒今朝醉，明日愁来明日愁。 | Wine brings Joy to ev'ry Heart, It lifts the Soul and cheers the Mind; From Sorrow's Grip it sets apart, And leaves our Worldly Cares behind. |
| <i>Modern</i>                            | 一杯美酒，带来欢乐。朋友相聚，笑语盈盈。             | A cup of wine brings joy. Friends gather, laughter fills the air.                                                                             |

Table 33: Representative samples from the contrastive style dataset. Examples preserve a common topic while varying historical register across four periods.

| English Query                                   | Chinese Query |
|-------------------------------------------------|---------------|
| Who are famous philosophers?                    | 著名的哲学家有哪些？    |
| What were major inventions?                     | 重大发明有哪些？      |
| What scripts were primarily used?               | 主要书写系统有哪些？    |
| How was time kept?                              | 记录时间的主要工具是什么？ |
| How did people communicate over long distances? | 主要远程通讯方式是什么？  |

Table 34: Sample queries from the Epistemic Cutoff Dataset. They test whether generated entities respect the knowledge boundary of the target era.

| English Query                                                           | Chinese Query         |
|-------------------------------------------------------------------------|-----------------------|
| I need to send a message from Beijing to London. How long will it take? | 我需要从北京给伦敦发送信息，需要多长时间？ |
| The night is dark. What do I do?                                        | 夜晚无灯，我该如何照亮？          |
| I want to preserve fish for the winter. What should I do?               | 我想保存鱼肉过冬，该怎么做？        |
| In winter it's cold. How do I keep the house warm?                      | 冬天寒冷，我如何使家中温暖？        |
| How can I cross a river?                                                | 要过河，应使用什么方式？          |

Table 35: Sample queries from the Causal Remodeling Dataset. They test whether outputs use technologies compatible with the target era.

| English Query                                                                                    | Chinese Query                  |
|--------------------------------------------------------------------------------------------------|--------------------------------|
| Please answer using modern English prose, what are the main methods for treating diseases today? | 请用现代汉语白话文体回答，当今时代治疗疾病的主要方法是什么？ |
| Please answer using modern English prose, what media are currently used to store information?    | 请用现代汉语白话文体回答，目前常用的信息存储介质是什么？   |
| Please answer using modern English prose, what sports are popular currently?                     | 请用现代汉语白话文体回答，目前流行的运动是什么？       |
| Please answer using modern English verse, what water supply systems exist in the modern era?     | 请用现代汉语诗歌文体回答，现代供水系统有哪些？        |
| Please answer using modern English verse, what painting materials were used in the present day?  | 请用现代汉语诗歌文体回答，现今绘画材料有哪些？        |

Table 36: Sample queries from the Mismatch Entanglement Dataset. Each query pairs a contemporary content or style request with historical steering.