HistLens: Mapping Idea Change across Concepts and Corpora

Yi Jing[♠◇*] **Weiyun Qiu**[♡*] **Yihang Peng**[♣] **Zhifang Sui**[◇†]
[♠]Department of Computer Science and Technology, Tsinghua University, China
[♡]School of History, Nanjing University, China
[♣]Department of Chinese Language and Literature, Tsinghua University, China
[◇]State Key Laboratory of Multimedia Information Processing,
School of Computer Science, Peking University, China
jingy22@mails.tsinghua.edu.cn, szf@pku.edu.cn

Abstract

Language change both reflects and shapes social processes, and the semantic evolution of foundational concepts provides a measurable trace of historical and social transformation. Despite recent advances in diachronic semantics and discourse analysis, existing computational approaches often (i) concentrate on a single concept or a single corpus, making findings difficult to compare across heterogeneous sources, and (ii) remain confined to surface lexical evidence, offering insufficient computational and interpretive granularity when concepts are expressed implicitly. We propose HistLens, a unified, SAE-based framework for multi-concept, multi-corpus conceptual-history analysis. The framework decomposes concept representations into interpretable features and tracks their activation dynamics over time and across sources, yielding comparable conceptual trajectories within a shared coordinate system. Experiments on long-span press corpora show that HistLens supports cross-concept, cross-corpus computation of patterns of idea evolution and enables implicit concept computation. By bridging conceptual modeling with interpretive needs, HistLens broadens the analytical perspectives and methodological repertoire available to social science and the humanities for diachronic text analysis.

Code LeoJ-xy/HistLens

1 Introduction

> *"Concepts are both indicators and factors of historical change."*
>
> —Reinhart Koselleck

Language and society co-evolve as two intertwined trajectories. Language change not only reflects but also shapes social processes, and the semantic evolution of fundamental concepts constitutes an observable marker of historical transformation. Concepts are more than lexical items: they are configurations of meaning, association, and argumentative roles embedded in social contexts, continuously reworked as social structures and public discourse shift. For computational social science and digital humanities, an important goal is therefore to characterize conceptual change in large-scale diachronic corpora, and to measure it in ways that support historical interpretation and social theory—ultimately yielding new, valuable insights for both social science and historiography.

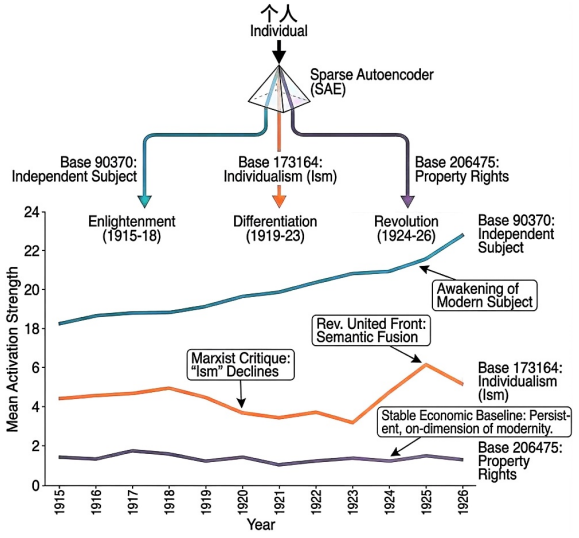

Figure 1: In *New Youth*, *individual* is not unitary but decomposes into sub-semantics such as *independent subjecthood*, *individualism as discourse*, and *property rights/economic individuality*, each with a partly independent trajectory over 1915–1926. HistLens identifies such internal conceptual heterogeneity and diachronic reconfiguration.

Recent years have witnessed substantial progress in NLP for diachronic semantics and discourse analysis, including work on lexical semantic change(Hamilton et al., 2016; Periti

*Equal contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14326–14351
July 2-7, 2026

and Tahmasebi, 2024), topic evolution(Blei and Lafferty, 2006; Dieng et al., 2019; Sirin and Lippincott, 2024; James et al., 2024), and the study of stances and frames in public discourse(Otmakhova et al., 2024; Irani et al., 2025). Yet, integrating these advances into a scalable, comparable, and interpretable paradigm for studying conceptual semantic evolution remains challenging.

First, **scalability and comparability across concepts and across corpora are still limited**. A large portion of existing work focuses on a single concept, a small set of keywords, or a single corpus(Tang et al., 2023; Gribomont, 2023; Matthews et al., 2025). Even when methods scale technically(James et al., 2024; Ma et al., 2025), their outputs are often not directly comparable across different concepts or across heterogeneous sources. This makes it difficult to address canonical questions such as: Do multiple concepts co-evolve in coordinated ways? Which changes are shared across corpora, and which are products of specific contexts?

Second, **existing approaches often fall short in characterizing implicit concepts**. Many methods remain centered on keywords and surface co-occurrence patterns, making it difficult to capture fine-grained intellectual and social evolution that is not explicitly expressed in canonical terms(Timkey and Schijndel, 2021; James et al., 2024; Otmakhova et al., 2024; Irani et al., 2025), thereby limiting analytical depth. For historians and social scientists, this yields two direct limitations: (i) "concept change" can be misread as mere lexical replacement or stylistic fluctuation, obscuring continuity and turns across shifting discursive strategies; and (ii) overlooking implicit expressions introduces bias in source selection and interpretation, so that identifying key turning points relies more on researchers' experience than on a systematic and objective evidential chain.

To address these challenges, we propose HistLens, a **unified conceptual-history analysis framework** for **multiple concepts** and **multiple corpora**. The framework is built on a sparse, feature-structured representational space: we decompose dynamic semantic representations into interpretable features, and recast conceptual inquiry as tracking the activation dynamics of these features across time and across sources. By anchoring different concepts within a shared sparse feature coordinate system, the framework enables consistent measurement and principled comparison across heterogeneous corpora. Empirically, we apply HistLens to long-span press corpora, demonstrating its scalability and its capacity to bridge computational modeling with humanistic interpretation.

HistLens is the first framework to enable unified measurement and comparable analysis of multiple concepts across multiple corpora within a shared, interpretable sparse feature space. This moves the study of concept evolution beyond keyword-level fluctuations by decomposing change into quantifiable and explainable semantic components. Moreover, it allows us to compute implicit semantic and ideational trajectories in texts where a concept is not explicitly named, offering a more systematic and interpretable quantitative methodology, and also a new analytic lens for humanistic close reading and social-scientific inquiry.

2 Related Works

2.1 Diachronic Semantic Change

A major computational route to conceptual history operationalizes "concept evolution" as diachronic semantic change, from long-run cultural and lexical trend quantification in massive digitized corpora (Michel et al., 2011; Hamilton et al., 2016) to contextual-representation pipelines that extract contextual word embeddings from masked language models and compare them across time via similarity or clustering (Devlin et al., 2019). Beyond term-level comparisons, related work models change via time-specific sense distributions (Tang et al., 2023), cross-temporal context perturbations (Aida and Bollegala, 2023), and multi-period diachronic sense induction (Periti and Tahmasebi, 2024); in digital humanities, semantic-difference keyword methods target cross-corpus divergence and interpretable "sites of semantic struggle" (Gribomont, 2023). Complementarily, concept/topic discovery examines how new semantic regions emerge and diffuse (Ma and Nyarko, 2025), while dynamic topic models (Blei and Lafferty, 2006), neural extensions such as the dynamic embedded topic model (Dieng et al., 2019), literary-historical operationalizations of that line of work (Sirin and Lippincott, 2024), and recent evaluations (James et al., 2024) support more reliable temporal topic comparisons.

2.2 Diachronic Modeling of Idea and Discourse

Beyond lexical meaning, computational social science and digital humanities study how stance, framing, and discourse regimes shift over time, with framing understood as a multi-dimensional construct involving selection, emphasis, narrative templates, and rhetoric rather than being reducible to topic or sentiment (Otmakhova et al., 2024). Related work quantifies longitudinal divergence such as polarization-driven drift in lexical choice, affect, and semantics (Karjus and Cuskley, 2024), models deliberation via computational argumentation and interaction dynamics (Irani et al., 2025), and uses cross-temporal semantic retrieval over news archives to trace narrative recurrence and transformation (Franklin et al., 2024); long-horizon historical corpora further motivate time-aware modeling of cultural change (Hegde et al., 2025; Ma et al., 2025).

2.3 Sparse Autoencoders for Interpretable Representations

A major interpretability approach decomposes dense neural representations into a small set of latent factors, mitigating multi-feature superposition (Elhage et al., 2022). Sparse coding and dictionary learning provide classical foundations (Olshausen and Field, 1996; Mairal et al., 2010), and sparse autoencoders instantiate these ideas for neural activations(Makhzani and Frey, 2013). In mechanistic interpretability, SAE-style dictionary learning has been used to extract more nearly monosemantic features from transformer activations and enable feature-level analyses at scale (Cunningham et al., 2023; Bricken et al., 2023; Gao et al., 2024), often paired with automated labeling and validation pipelines (Bills et al., 2023; O'Neill et al., 2024). By contrast, direct comparisons of word vectors can be affected by anisotropy, rogue dimensions, robustness issues, and social bias (Ethayarajh, 2019; Timkey and Schijndel, 2021; Guo and Caliskan, 2021), motivating sparse-feature-based analyses for more robust characterization of linguistic and semantic phenomena (Matthews et al., 2025; Jing et al., 2025).

3 Methodology

3.1 Sparse Auto-Encoder Representations

We use pretrained Sparse Auto-Encoders (SAEs) as a fixed system for diachronic analysis. Given timestamped textual units $\{(x_i, t_i)\}$, where each x_i is a sentence, we encode each x_i with a frozen pretrained LLM to obtain token-level hidden states $\mathbf{h}_{i,j} \in \mathbb{R}^d$. For the main pipeline, $\mathbf{h}_{i,j}$ denotes the residual-stream state at Layer 29 of Llama-3.1-8B-Instruct, and we use the pretrained Layer-29 OpenSAE as a fixed mapping throughout unless otherwise noted; additional technical background is given in Appendix A. A pretrained SAE then maps each hidden state into a sparse feature space,

$$\mathbf{z}_{i,j} = f_{\mathrm{SAE}}(\mathbf{h}_{i,j}) \in \mathbb{R}^K, \tag{1}$$

and we aggregate token activations into a text-level sparse representation $\mathbf{z}_i = \mathrm{Agg}_j(\mathbf{z}_{i,j})$ using max pooling over tokens. Throughout, we *do not* update parameters of either the LLM or the SAE; we treat them as fixed nonlinear mappings and analyze only the resulting activations.

Selecting drifting base vectors and interpretation. For each SAE base vector (dimension) k, we summarize its activation by time slice s as $\mu_{k,s} = \mathbb{E}[\mathbf{z}_{i,k} \mid t_i \in s]$, and define its cumulative drift over a given period as

$$D_k = \sum_{s=2}^{S} |\mu_{k,s} - \mu_{k,s-1}| \,. \tag{2}$$

We then select base vectors with the largest D_k (i.e., maximal total drift). To interpret each selected vector, we retrieve its highest-activating texts and submit these contexts to human experts, who assign semantic descriptions grounded in the evidence provided by the activating passages.

3.2 Diachronic Analysis

HistLens treats conceptual history as the reconfiguration of *interpretable semantic components* across time and discourse fields, operationalized as SAE *base vectors* in a shared sparse activation space (Section 3.1). Each concept is represented as a structured combination of base-vector activations, enabling scalable distant reading with evidential accountability: quantitative signals locate salient periods and components, and claims are traceable to high-activating textual evidence.

3.2.1 Constructing the Concept–Corpus Atlas

We slice each periodical into yearly bins and, for each (concept, corpus) pair, compute a small set of

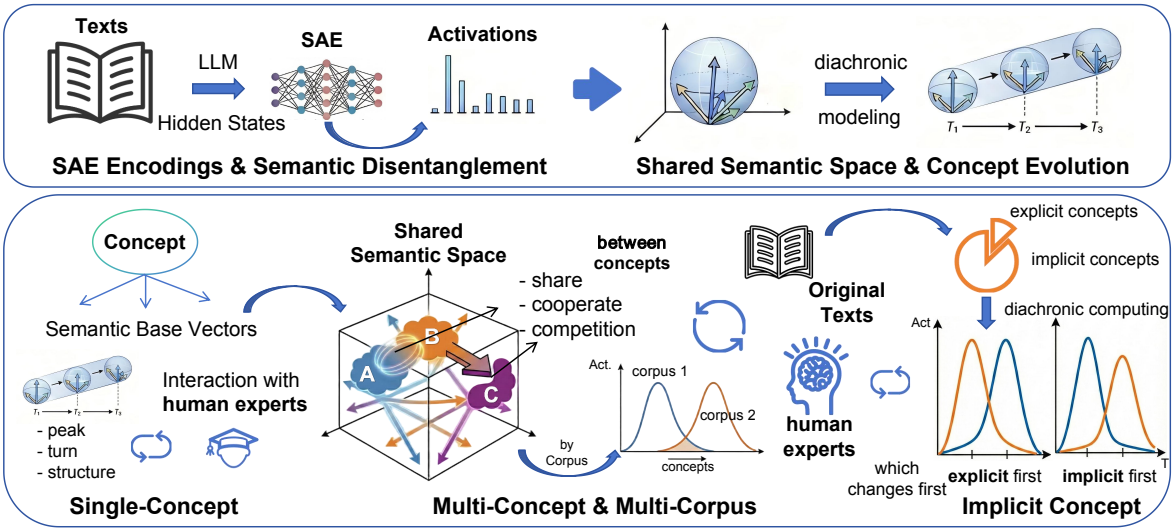

Figure 2: **Framework overview**. We encode diachronic texts with an LLM and project hidden states into sparse SAE activations to obtain interpretable semantic base vectors, forming a shared semantic space. We then perform diachronic modeling in this space to quantify concept dynamics (e.g., peak windows, turning points, and structural shifts) for single-concept and multi-concept/multi-corpus settings, and further distinguish implicit concept signals from explicit lexical mentions, offering new insights for the study of intellectual and conceptual evolution.

navigational statistics to enable macro characterization and cross-material comparison (formal definitions in Appendix B). The *peak year* is the year where the concept magnitude (slice-aggregated activation over $\mathcal{S}_c$) is maximal. The *turning point* is the year with the strongest adjacent-slice change; its signed intensity I equals the signed magnitude of that change (positive for increase, negative for decrease). The *diversity* H quantifies how dispersed the concept composition is across base vectors, defined as the normalized entropy of base-vector contribution shares (higher H indicates a more diffuse mixture).

3.2.2 Single-Concept Diachronic Decomposition

For a single concept c, we decompose its diachronic variation in the shared SAE space into a small set of informative base-vector drivers. For each time slice s, we compute the mean activation of base vector k as $\mu_{k,s} = \mathbb{E}[z_{i,k} \mid t_i \in s]$, and select a concise set of base vectors with high activation in texts salient for concept c and salient temporal variation. We then track $\mu_{k,s}$ trajectories over time and use adjacent-slice differences (or relative change rates) to locate localized surges and reversals, yielding fine-grained quantitative signals that guide subsequent analysis.

3.2.3 Multi-Concept Comparability and Interaction

Because all concepts are embedded in the same base-vector coordinate system, comparisons across concepts do not require retraining concept-specific spaces. We therefore analyze multi-concept dynamics at the level of (i) orientation trajectories and (ii) shared base-vector signals, which preserves comparability while keeping interpretation tethered to evidence. Methodologically, this supports humanities questions about conceptual co-evolution, alignment, and tension as relations among interpretable components, rather than as opaque movements in embedding space. The corresponding formal quantities used for multi-concept comparison are specified in Appendix B.

3.2.4 Cross-Corpus Comparability

Let $\mathcal{R}$ denote a set of corpora (e.g., periodicals or genres). Cross-corpus analysis conditions the same conceptual quantities on $r_i \in \mathcal{R}$ and computes corpus-aware summaries within each time slice. To keep heterogeneous corpora comparable, we focus on *salient realizations* of each concept within each corpus: for each (c, r) we select a high-quantile subset $\mathcal{I}_{c,r}$ of text units by concept magnitude and summarize diachronic coverage and composition only within this set (formal definitions in Appendix B). Since all corpora share the same

SAE space, we further align *semantic components* across corpora by comparing the overlap of Top-30 drifting base vectors, separating shared components from corpus-specific reweighting.

3.2.5 Implicit Concept Computation and Interpretation

Concepts are often expressed without their canonical names, via stable discursive patterns. We therefore split the salient set $\mathcal{I}_{c,r}$ into *lexically anchored* vs. *implicit* contexts by whether canonical lexemes of concept c are present (see Appendix B.5 for details), and let $\mathcal{I}^{\text{Imp}}_{c,r}$ be the implicit subset. We quantify implicit realization by the ratio

$$\bar{r}_{c,r} = \frac{\sum_{i \in \mathcal{I}^{\text{Imp}}_{c,r}} m_i^{(c)}}{\sum_{i \in \mathcal{I}_{c,r}} m_i^{(c)}}. \tag{3}$$

Humanistic interpretation. We treat computational signals as methodological scaffolding and follow a humanistic interpretation protocol based on expert reading of highest-activating contexts (Appendix B.9).

4 Experiments

4.1 Experiment Setup

We conduct our experiments with the pretrained OpenSAE family (THU-KEG, 2025) attached to Llama-3.1-8B-Instruct (Grattafiori et al., 2024). Unless otherwise stated, all main reported results use the Layer-29 SAE over the residual stream; only the dedicated cross-layer robustness analysis reruns the same pipeline with Layers 06/14/22/29. We compile and curate complete runs of several representative periodicals from modern Chinese history, including *New Youth (Xinqingnian)* and *The Guide (Xiangdao)*, digitized from print editions via OCR, totaling *3,277* issues and *8,030,009* characters. We further select four foundational concepts as case studies: *individual*, *society*, *nation*, *world*.

4.2 Main Results

4.2.1 Concept–Corpus Atlas

We slice each corpus into time-indexed segments and pass them through the SAE to obtain sparse-feature activations. Following Section 3.1, we compute a compact set of reproducible statistics for each (concept, corpus) pair: de-lexicalization strength $\bar{r}$, diversity H, the peak year, and the strongest turning point (Table 1).

Concept	$\bar{r}$	H	Peak	Turn (year, I)
New Youth				
individual	0.920	0.741	1920	(1918, +0.226)
nation	0.921	0.743	1924	(1918, +0.116)
society	0.595	0.368	1922	(1918, −0.213)
world	0.900	0.683	1926	(1918, +0.230)
The Guide				
individual	0.963	0.763	1923	(1923, +0.0665)
nation	0.860	0.568	1926	(1923, −0.0810)
society	0.759	0.557	1924	(1924, +0.152)
world	0.778	0.786	1925	(1926, −0.0852)

Table 1: A slice of the Concept–Corpus Atlas. We report ratio-based and structural signals only: de-lexicalization strength $\bar{r}$, diversity H (normalized entropy over orientations), and diachronic anchors (Peak year and the strongest turning point). Turn intensity I is signed: $I > 0$ indicates an upward shift and $I < 0$ a downward shift at the turning year.

Overall, de-lexicalized concept practice is substantial, with $\bar{r}$ ranging from 0.595 to 0.963. In *New Youth*, *individual*, *nation*, and *world* are strongly de-lexicalized ($\bar{r} \approx 0.900$–0.921), whereas *society* is less so ($\bar{r} = 0.595$), indicating more lexically anchored usage. Structural dispersion also varies: *society* in *New Youth* is markedly concentrated ($H = 0.368$), while *world* in *The Guide* is more diffuse ($H = 0.786$). Diachronic anchors sharply separate the corpora: all four concepts in *New Youth* share their strongest turn at 1918 with larger magnitudes ($|I| = 0.116$–0.230), whereas the strongest turns in *The Guide* shift later (1923–1926) and are weaker overall (largest $|I| = 0.152$). These compact signals provide reproducible anchors for selecting downstream case studies in the subsequent analyses.

4.2.2 Single-Concept Analysis

Guided by the *Concept–Corpus Atlas*, we zoom in on a single concept, *individual*, in *New Youth* to illustrate the value of SAE-based decomposition: an apparently holistic diachronic trajectory can be factorized into a small number of interpretable semantic drivers. We select several highly activated base vectors with strong explanatory power and track their mean activations across time slices. To reduce cognitive load, we refer to these components below by short semantic labels rather than by raw base indices; Table 6 maps each label to its constituent SAE base(s) and representative bilingual evidence.

As shown in Fig. 3, the evolution of *individual*

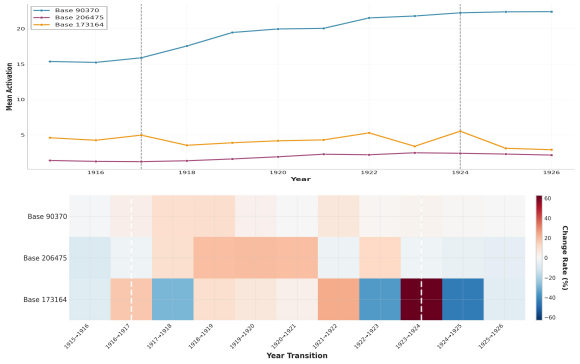

Figure 3: Single-concept decomposition for *individual* in *New Youth*. Top: mean activation trajectories of selected SAE base vectors across years, with dashed markers indicating salient transition windows. Bottom: year-to-year relative change rates for the same base vectors, highlighting localized surges and reversals that serve as navigational anchors for close reading.

decomposes into three primary semantic strands: *Actorhood* rises steadily over the long run and remains high in later years; *Individualism as Discourse* exhibits a sharp surge in a salient turning window followed by a rapid retreat; and *Property and Economic Individuality* stays comparatively low but increases gradually, contributing as a persistent background dimension. Overall, the concept's change is driven by differentiated dynamics at the semantic-component level rather than incidental fluctuations in surface frequency or collocational patterns.

This decomposition further indicates that, in this diachronic corpus, *individual* is not a semantically homogeneous object that can be adequately captured by a single scalar indicator; instead, it forms a "semantic assemblage" composed of multiple strands that can vary relatively independently (e.g., actorhood, discursive individualism, and property/economic individuality). Accordingly, seemingly paradoxical historical observations—for example, periods in which *individualism* as discourse recedes while discussion of *individual* does not diminish—need not be interpreted as conceptual disappearance or analytic misreading. Rather, they reflect shifts in the relative prominence and coordination among internal strands: conceptual continuity derives from the persistence of the assemblage, whereas conceptual turning points emerge from reorganization within the assemblage. The SAE-based framework thus provides a principled lens for characterizing the internal heterogeneity of concepts and tracing its diachronic evolution.

4.2.3 Multi-concept analysis

We conduct an aligned comparison of four concepts in *New Youth*—*individual*, *society*, *nation*, and *world*—within a shared SAE semantic coordinate system. Concretely, we select a small set of representative sparse features for each concept, compute their mean activations on yearly slices, and aggregate feature shares at the window level. We then use difference heatmaps to contrast within-concept reweighting in 1917–1919 relative to *pre*1917 and in 1922–1924 relative to 1917–1919, thereby characterizing multi-concept dynamics on the same timeline and in the same semantic units. Throughout this discussion, we refer to the selected features by stable semantic labels; Table 6 provides the label-to-base mapping.

As visualized in Fig. 4, the difference patterns indicate structured semantic reconfiguration around major historical junctures. In 1917–1919 relative to *pre*1917, *individual* increases *Actorhood* while decreasing *Individualism as Discourse*; in parallel, *society* exhibits a relative rise in *Societal Transition and Institutional Design*, accompanied by relative declines in *Organized Praxis and Labor-Movement Alignment* and *Party Linkage and Organizational Alignment*. Moving to 1922–1924, reweighting further intensifies: *society* shifts from *Societal Transition and Institutional Design* toward *Organized Praxis and Labor-Movement Alignment*; *world* strengthens *Revolutionary International Field* while *Actorhood* relatively decreases; and *nation* increases *Nation-State as Strategic Instrument*, reflecting a temporally localized restructuring of the national frame.

These shifts suggest that the central object of concept computation is not the mere appearance or disappearance of a concept, but the reweighting of its internal semantic components under changing historical pressures, which in turn reshapes its discursive function and historical role. *Individual* becomes more strongly organized around narratives of agentive position and responsibility; *society* reallocates emphasis between institutional blueprints and organized praxis; *world* is more tightly embedded in transnational revolutionary linkages; and *nation* is foregrounded as a strategic instrument and a structure of contestation. Building on this view, we operationalize the otherwise hard-to-systematize notion of "shifts in semantic emphasis" as fine-grained reweighting and cross-concept alignment in a shared SAE space, opening

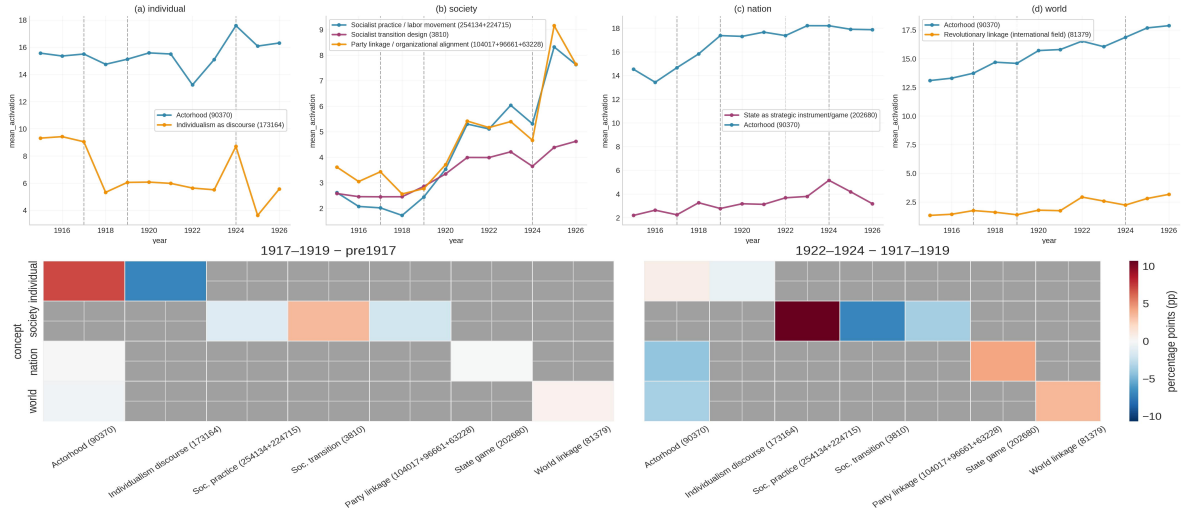

Figure 4: Multi-concept dynamics in a shared SAE space. Top: for *individual*, *society*, *nation*, and *world*, we track a small set of salient sparse features and plot their mean activations by year, enabling direct comparison in the same semantic coordinate system. Bottom: window-level composition shifts are summarized as difference heatmaps of feature-share changes across two contrasts, 1917–1919 relative to *pre*1917 and 1922–1924 relative to 1917–1919, revealing coordinated within-concept reweighting around historical junctures.

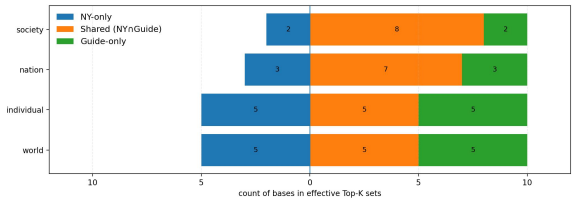

Figure 5: Cross-corpus decomposition of effective Top-30 drifting bases in a shared SAE space. For each concept, we partition the drifting bases into *New Youth*-only, shared (*New Youth* ∩ *The Guide*), and *The Guide*-only components, visualizing how cross-corpus overlap versus corpus-specific components vary across concepts.

up additional quantitative entry points while preserving the nuance of humanities interpretation.

4.2.4 Cross-corpus Analysis

We further conduct a cross-corpus comparison between *New Youth* and *The Guide*. Within the same SAE semantic coordinate system, we first quantify the overlap of the Top-30 drifting bases to distinguish a cross-corpus *stable semantic skeleton* from *corpus-specific semantic reweighting*. We then use, for each base, its full-range top-30 highest-activating sentences as the evidence pool and display 8 representative sentences for close reading, attributing divergences to concrete discourse mechanisms (e.g., genre-specific communicative tasks, polemical targets, and mobilization-oriented rhetoric). This workflow makes cross-material comparison less dependent on ad hoc sentence selection: instead, comparable semantic components provide a reproducible reading guide, enabling more transparent and falsifiable historical interpretation.

We illustrate the value of this multi-source perspective with the concept *world*. Its relatively low overlap in Top-30 drifting bases suggests that *world* is not organized by an identical set of semantic handles across the two corpora; nevertheless, the shared drifting bases still provide a common baseline, consistently anchoring *world* to a macro-structure of revolution, class struggle, and imperialism. The humanities-oriented gain lies in the *localizable* nature of the differences: in *New Youth*, *world* more readily intertwines with intellectual debate, epistemic framing, and cultural critique, forming a world-view oriented toward conceptual renovation; in *The Guide*, *world* is more often compressed into a field of camp alignment, organizational mobilization, and institutional claims, yielding a sharper action orientation and political demarcation. In this sense, cross-corpus comparison turns the heterogeneity of *world* into a structural finding: the same-named concept is reweighted by different discursive tasks on top of a shared backbone, resulting in distinct rhetorical foci and historical functions. Our innovation is to operationalize cross-corpus comparability as a decomposition into shared versus corpus-specific semantic components with aligned evidence, thereby combining the nuance of close reading with a scalable and re-

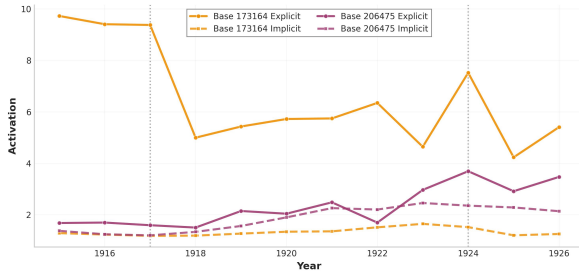

Figure 6: Explicit vs. implicit concept computation for *individual* in *New Youth*. Yearly mean activations are shown for *Individualism as Discourse* and *Property and Economic Individuality* in anchor-present (Explicit) and anchor-absent (Implicit) contexts. Vertical dashed lines indicate salient transition windows.

Concept	Layer	Peak	Turn	Avg. Jaccard
individual	06	1923	1918	0.50
	14	1923	1918	0.42
	22	1920	1918	0.48
	29	1920	1918	0.33
nation	06	1924	1918	0.57
	14	1924	1918	0.53
	22	1924	1918	0.60
	29	1924	1918	0.57

Table 2: Cross-layer robustness results on *New Youth*. "Avg. 2-gram Jaccard" is the mean 2-gram Jaccard similarity between the evidence-context fingerprint of a given layer and those of the other three layers.

producible computational procedure.

4.2.5 Implicit Concept Computation

We further contrast *Explicit* (anchor-present) and *Implicit* (anchor-absent) evidence for *individual* in *New Youth* by tracking yearly mean activations of two representative semantic components, *Individualism as Discourse* and *Property and Economic Individuality* (Fig. 6). The results show that concept-related semantics remains substantial even when lexical anchors are absent, and that implicit dynamics are not a trivial mirror of explicit trajectories, providing additional and reproducible navigational signals for close reading. A separate historian validation on 20 sampled anchor-absent high-activation cases per concept yields semantic-consistency rates of 90% for *society*, 100% for *nation*, 100% for *world*, and 75% for *individual*; representative examples are reported in Appendix C.

For *Individualism as Discourse*, the Explicit trajectory is high in 1915–1917, drops sharply in 1918, and rebounds during 1922–1924 with a local peak around 1924. In contrast, the Implicit trajectory stays lower and smoother, reaching a relative high around 1923 before retreating. This divergence suggests that explicit articulation is more sensitive to episodic contestation and rhetorical intensity, whereas anchor-absent semantic practice is more stable, indicating persistence of the concept beyond overt lexical naming.

For *Property and Economic Individuality*, Implicit evidence rises steadily from 1919 and forms a plateau around 1922–1923, while the Explicit curve exhibits a pronounced jump only in 1923–1924 and peaks around 1924, displaying an implicit-leads-explicit pattern. This implies that institutional and economic semantic components may diffuse via stable semantic carriers across heterogeneous topics before becoming explicitly foregrounded and lexically consolidated in more concentrated political and institutional debates.

Overall, Explicit signals are closer to public naming and the intensity of contestation, whereas Implicit concept computation better captures semantic permeation and cross-topic diffusion. Their divergence and lead–lag relations serve as key criteria for locating turning windows for downstream close reading. This explicit–implicit contrastive quantification framework renders the *presence and change of a concept even when its keyword is not explicitly attested in the text* into computable and interpretable outcomes, advancing the traditionally intuition-driven identification of unspoken ideas into a systematic line of inquiry.

4.2.6 Cross-layer Analysis

To test whether our diachronic signals depend on the SAE layer, we rerun the same pipeline on the same corpus while only switching SAE layers (06/14/22/29), and use *individual* and *nation* in *New Youth* as representative concepts. For each layer, we report (i) the *peak year*, defined as the year in which the concept magnitude attains its maximum; (ii) the *turning year*, defined as the year corresponding to the strongest adjacent-slice change; and (iii) *Avg. Jaccard*, the mean 2-gram Jaccard similarity between the evidence-context fingerprint of a given layer and those of the other three layers (Table 2). As shown in Table 2, the turning year is identical across all four layers for both concepts (1918 throughout), indicating that the turning-point signal used to locate major semantic reconfiguration is insensitive to layer choice and can serve as a reproducible temporal anchor.

Meanwhile, peak-year localization exhibits concept-dependent layer sensitivity. For *nation*, the peak year is fully stable (1924 across all layers), suggesting that the intensity-based localization of *nation* is robust to representational depth in our setup. In contrast, *individual* shows a consistent stratification: lower layers (06/14) peak in 1923, whereas higher layers (22/29) peak earlier in 1920. This pattern suggests that, while the major turning structure is preserved across layers, different representational depths may emphasize different facets of the concept and thus shift where the maximal concentration of evidence is localized.

This layer effect is also reflected in contextual consistency. Using evidence-context 2-gram fingerprints, *nation* exhibits higher and relatively stable cross-layer agreement (Avg. Jaccard ≈ 0.53–0.60), whereas *individual* is less consistent overall and becomes notably more divergent at the highest layer (Layer 29: Avg. Jaccard $= 0.33$). Overall, the cross-layer experiment supports a robust diachronic backbone (the 1918 turning point) while indicating that peak localization and the concrete evidential contexts can vary more for *individual* than for *nation*, providing complementary views on how conceptual signals are distributed across representational depth.

5 Conclusion

We propose HistLens, a unified computational framework for conceptual history: within a shared, alignable SAE semantic space, it jointly characterizes and compares the diachronic dynamics of multiple concepts across heterogeneous sources. We further introduce *implicit concept computation*, which captures stable semantic signals even when a concept is not explicitly stated by its canonical keywords. Overall, HistLens translates conceptual change in historical texts into an interpretable, comparable, and traceable quantitative structure, offering the humanities and social sciences new tools and perspectives for diachronic text analysis.

6 Limitations

Our work has limitations in experimental scale and in the monosemy (semantic univocality) of SAE base vectors.

Experimental scale. Because manual curation and interpretation of both corpora and model outputs is labor-intensive, we conduct case studies on only a limited set of periodicals and a limited set of concepts. In principle, the proposed framework has clear potential to scale to much larger collections. Future work may leverage LLM-based agents to assist analysis and synthesis; however, we remain cautious about the reliability of conclusions produced by current LLMs when applied to large-scale humanities and social-science analysis.

Monosemy of base vectors. Due to training-cost constraints, the SAE base vectors we obtain are not uniformly monosemous with a clear semantic referent. In our experiments, we observe that some base vectors do not exhibit a stable single meaning, and we exclude them from downstream analysis. Addressing monosemy in practical SAE deployments remains a major open challenge for the field.

7 Ethical Considerations

This section discusses the ethical considerations and broader impact of this work:

Potential Risks: Our framework provides structured access to latent semantic features and their diachronic dynamics. While the goal is scholarly analysis and evidence-grounded interpretation, similar tooling could be repurposed to engineer persuasive narratives or selectively amplify particular framings in downstream generation settings. To mitigate misuse, we will release code and documentation with clear intended-use statements, emphasize analysis over manipulation, and provide reproducible evaluation scripts so that claims can be independently verified.

Intellectual Property: The models and toolkits used in this study, including Llama-3.1-8B(-Instruct) and the OpenSAE framework, are open-source and used for scientific research in accordance with their respective licenses. We will also document provenance for all corpora and derived artifacts, and follow any redistribution constraints associated with the underlying sources.

Data Privacy: Our experiments use historical periodicals from the public domain or with established research access. The data does not target individuals and is not intended to contain personal or private information. We additionally apply basic filtering to remove incidental personal identifiers that may appear in historical texts where applicable.

Intended Use: HistLens is intended as a research tool for computational social science and digital humanities: it supports scalable concept analysis, facilitates evidence retrieval for close reading, and enables reproducible measurement of conceptual reconfiguration over time and across sources. It is not intended for decision-making about individuals or for high-stakes profiling.

Documentation of Artifacts: We will comprehensively document all released artifacts (corpora metadata, preprocessing, slicing procedures, probe construction, and evaluation metrics), including their domains, time spans, languages, and known limitations, to ensure transparency and reproducibility.

AI Assistants in Research or Writing: We employ Cursor for code development assistance and use GPT-5.2 to refine and polish the language of the manuscript.

Acknowledgments

This work is supported by NSFC Project 62476009, the Open Project Fund of the State Key Laboratory of Multimedia Information Processing (Project No. SKLMIP-KF-2025-01), and the Tsinghua University Disruptive Innovation Talent Development Program.

References

Taichi Aida and Danushka Bollegala. 2023. Swap and predict – predicting the semantic changes in words across corpora by context swapping. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7753–7772, Singapore. Association for Computational Linguistics.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 113–120. ACM.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Published Oct 4, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Hongmin Chen and Bingbing Wei. 2005. A preliminary study of the chinese communist party's propaganda strategy during the national revolution: Centered on *The Guide*, 1923–1925. *Anhui Historiography*, (4):63–70. In Chinese.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert P. Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. arXiv preprint. *Preprint*, arXiv:2309.08600.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. The dynamic embedded topic model. arXiv preprint. *Preprint*, arXiv:1907.05545.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. arXiv preprint. *Preprint*, arXiv:2209.10652.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65. Association for Computational Linguistics.

Brevin Franklin, Emily Silcock, Abhishek Arora, Tom Bryan, and Melissa Dell. 2024. News deja vu: Connecting past and present with semantic search. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 99–112, Mexico City, Mexico. Association for Computational Linguistics.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. arXiv preprint. *Preprint*, arXiv:2406.04093.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Isabelle Gribomont. 2023. From diachronic to contextual lexical semantic change: Introducing semantic difference keywords (sdks) for discourse studies. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 153–160, Singapore. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Niharika Hegde, Subarnaduti Paul, Lars-Joel Frey, Manuel Brack, Kristian Kersting, Martin Mundt, and Patrick Schramowski. 2025. Chronoberg: Capturing language evolution and temporal awareness in foundation models. arXiv preprint. *Preprint*, arXiv:2509.22360.

Matthew D. Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23*, pages 856–864.

Arman Irani, Ju Yeon Park, Kevin M. Esterling, and Michalis Faloutsos. 2025. A discourse analysis framework for legislative and social media debates. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 199–209, New Brunswick, NJ, USA. Association for Computing Machinery.

Charu Karakkaparambil James, Mayank Nagda, Nooshin Haji Ghassemi, Marius Kloft, and Sophie Fellenz. 2024. Evaluating dynamic topic models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 160–176, Bangkok, Thailand. Association for Computational Linguistics.

Yi Jing, Zijun Yao, Hongzhu Guo, Lingxu Ran, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2025. Lingualens: Towards interpreting linguistic mechanisms of large language models via sparse auto-encoder. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28232–28251, Suzhou, China. Association for Computational Linguistics.

Donald A. Jordan. 2019. *The Northern Expedition: China's National Revolution of 1926–1928*. University of Hawaii Press, Honolulu.

Andres Karjus and Christine Cuskley. 2024. Evolving linguistic divergence on polarizing social media. *Humanities and Social Sciences Communications*, 11(1):1–14.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Sibo Ma and Julian Nyarko. 2025. Identifying emerging concepts in large corpora. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6760–6778, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuxi Ma, Yongqian Peng, and Yixin Zhu. 2025. Word embeddings track social group changes across 70 years in china. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2010. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60.

Alireza Makhzani and Brendan J. Frey. 2013. k-sparse autoencoders. arXiv preprint. *Preprint*, arXiv:1312.5663.

Jacob A. Matthews, Laurent Dubreuil, Imane Terhmina, Yunci Sun, Matthew Wilkens, and Marten Van Schijndel. 2025. Disentangling language change: sparse autoencoders quantify the semantic evolution of indigeneity in french. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11208–11222, Albuquerque, New Mexico. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Yannian Ni. 2021. The three transformations of *New Youth* and the marker of the origin of the communist party press enterprise. *Modern Communication*, (8):29–35. In Chinese.

Bruno A. Olshausen and David J. Field. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407–15428, Bangkok, Thailand. Association for Computational Linguistics.

Charles O'Neill, Christine Ye, Kartheik G. Iyer, and John F. Wu. 2024. Disentangling dense embeddings

with sparse autoencoders. arXiv preprint. *Preprint*, arXiv:2408.00657.

Francesco Periti and Nina Tahmasebi. 2024. Towards a complete solution to lexical semantic change: an extension to multiple time periods and diachronic word sense induction. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 108–119, Bangkok, Thailand. Association for Computational Linguistics.

Hale Sirin and Tom Lippincott. 2024. Dynamic embedded topic models and change-point detection for exploring literary-historical hypotheses. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 231–236, St. Julians, Malta. Association for Computational Linguistics.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. A word sense distribution-based approach for semantic change prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.

THU-KEG. 2025. Opensae: Open-sourced sparse autoencoder towards interpreting large language models. `https://github.com/THU-KEG/OpenSAE`.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. arXiv preprint. *Preprint*, arXiv:2109.04404.

Patricia Uberoi. 1972. Nationalism and internationalism: Conflicting values of the chinese new culture movement. *China Report*, 8(1–2):54–62.

C. Martin Wilbur. 1984. *The Nationalist Revolution in China, 1923–1928*. Cambridge University Press, Cambridge.

A OpenSAE Technical Background

Architecture and placement. OpenSAE (THU-KEG, 2025) is a family of pretrained sparse autoencoders released for LLaMA-3.1-8B. The released SAEs are pretrained on the residual stream, use a context length of 4096, are trained on 22B tokens, and expand the LLaMA-3.1-8B hidden state by a factor of $64\times$ into a feature space of size 262,144. In this paper, unless otherwise noted, the main pipeline uses the checkpoint attached to Layer 29. Accordingly, we write $\mathbf{h}_{i,j}^{(29)} \in \mathbb{R}^d$ for the Layer-29 residual-stream state of token j in text unit i.

Encoder, sparse activation, and decoder. OpenSAE maps each layer-specific hidden state to a high-dimensional sparse feature space. Let $\tilde{\mathbf{h}}_{i,j}^{(29)}$ denote the normalized input passed to the encoder. The dense pre-activation vector is

$$\mathbf{a}_{i,j} = W_{\text{enc}}\tilde{\mathbf{h}}_{i,j}^{(29)} + \mathbf{b}_{\text{enc}} \in \mathbb{R}^K. \tag{4}$$

Here $K = 262{,}144$ is the total feature dimension of the released OpenSAE checkpoint family used in our study. OpenSAE then applies a TopK sparsification operator that retains only a small set of activated coordinates:

$$[\mathbf{z}_{i,j}]_m = \begin{cases} [\mathbf{a}_{i,j}]_m, & m \in \text{TopK}_\kappa(\mathbf{a}_{i,j}), \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where $\text{TopK}_\kappa(\mathbf{a}_{i,j})$ denotes the index set of the retained features and κ is the number of active features kept for each token. The decoder reconstructs the residual-stream state as

$$\hat{\mathbf{h}}_{i,j}^{(29)} = W_{\text{dec}}\mathbf{z}_{i,j} + \mathbf{b}_{\text{dec}}. \tag{6}$$

The sparse activation vector $\mathbf{z}_{i,j}$ is the object used throughout our diachronic pipeline.

How the SAE outputs are interpreted. The primary output used in our analysis is the sparse activation vector $\mathbf{z}_{i,j}$. An active coordinate $[\mathbf{z}_{i,j}]_m > 0$ indicates that the token-level residual state $\mathbf{h}_{i,j}^{(29)}$ aligns with feature m strongly enough to survive the TopK selection step, while its magnitude indicates the relative strength of that alignment for the given token. We therefore interpret the nonzero support of $\mathbf{z}_{i,j}$ as a sparse inventory of semantic components expressed by the token in context. The associated decoder direction, given by the m-th column of W_{dec} up to implementation convention, specifies how that feature contributes back to the reconstructed hidden state and thus provides the geometric direction that the feature adds to the residual stream.

This interpretation is evidential rather than purely label-based. A feature is not treated as meaningful merely because it has a nonzero index; instead, its interpretation is established by inspecting its highest-activating contexts, verifying that those contexts exhibit a stable semantic regularity, and then assigning a description grounded in that regularity. At the text level, aggregated activations $\mathbf{z}_i$ indicate the relative prevalence of interpretable semantic components within a sentence, rather than the presence of a single categorical meaning. In the main pipeline, $\mathbf{z}_i$ is obtained by max pooling token-level SAE activations within the sentence. By contrast, the reconstruction $\hat{\mathbf{h}}_{i,j}^{(29)}$ is not itself read as a human-interpretable semantic object; it is used as evidence that the retained sparse features preserve substantial information about the original residual-stream state.

Training objective. At the level documented by the public OpenSAE interface, the training loss can be written as

$$\begin{aligned}\mathcal{L}_{\text{SAE}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{AuxK}}\mathcal{L}_{\text{AuxK}} \\ + \lambda_{\text{MTK}}\mathcal{L}_{\text{MultiTopK}} + \lambda_1\mathcal{L}_1,\end{aligned} \tag{7}$$

where the reconstruction term is the squared L_2 loss between the input hidden state and its reconstruction,

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{i,j}\left[\left\|\hat{\mathbf{h}}_{i,j}^{(29)} - \mathbf{h}_{i,j}^{(29)}\right\|_2^2\right]. \tag{8}$$

The remaining terms correspond to the AuxK loss, the Multi-TopK loss, and the L_1 regularization term exposed by OpenSAE (THU-KEG, 2025). Since our study uses released pretrained checkpoints rather than retraining the SAE, these auxiliary terms matter only as part of the pretrained representation family from which our sparse activations are drawn.

Interpretive role in this paper. In our use case, both the LLM and the SAE remain frozen. We do not optimize the SAE further and do not alter its training objective. Instead, we treat the learned sparse features and their associated decoder directions as reusable semantic components, and conduct all downstream analysis on the activations $\mathbf{z}_{i,j}$ and their text-level aggregations $\mathbf{z}_i$. This fixed-feature regime is what enables the subsequent anal-

yses of drift, cross-corpus overlap, implicit realization, and evidence-grounded interpretation.

B Diachronic Computations and Definitions

B.1 Notation and Time Slicing

Time slicing and conditioning. Let $t \in \{1, \ldots, T\}$ index ordered time slices. We write $r \in \mathcal{R}$ for corpora/sources and i for sentence-level text units with metadata (t_i, r_i) and SAE activations $\mathbf{z}_i$.

Concept-level quantities and layer convention. Unless otherwise noted, all quantities in this appendix are computed from the Layer 29 residual-stream OpenSAE used in the main pipeline, so $\mathbf{z}_i$ denotes the sentence-level max-pooled aggregation of Layer-29 token activations. For each concept c, we instantiate a fixed, study-specific set $\mathcal{S}_c$ of expert-validated SAE bases (or base clusters). $\mathcal{S}_c$ is treated as an operational concept definition for downstream diachronic computation rather than as the output of a separate automatic learning algorithm. For a text unit i and component $s \in \mathcal{S}_c$, let $Z_i^{(c,s)} \geq 0$ denote the aggregated activation assigned to s. We define the concept magnitude of text unit i by

$$m_i^{(c)} = \sum_{s \in \mathcal{S}_c} Z_i^{(c,s)}. \tag{9}$$

B.2 Salient Set Construction

Within-corpus salience threshold. For each (c, r), define a high-quantile threshold over concept magnitudes within corpus r:

$$\tau_{c,r} = \mathrm{Quantile}_q\Big(\{\, m_i^{(c)} \mid r_i = r \,\}\Big), \tag{10}$$

where $q = 0.95$ in all reported experiments.

Salient realizations. The salient set used for corpus-aware diachronic summaries is

$$\mathcal{I}_{c,r} = \{\, i : r_i = r,\ m_i^{(c)} \geq \tau_{c,r} \,\}. \tag{11}$$

B.3 Aggregation and Composition

Within-corpus, within-time aggregation. Define within-corpus, within-time means on the salient set:

$$\mu_{c,s,r,t} = \mathbb{E}\Big[Z_i^{(c,s)} \;\Big|\; i \in \mathcal{I}_{c,r},\ t_i = t \Big], \tag{12}$$

$$A_{c,r,t} = \mathbb{E}\Big[m_i^{(c)} \;\Big|\; i \in \mathcal{I}_{c,r},\ t_i = t \Big]. \tag{13}$$

Orientation shares (composition). Orientation shares are computed as

$$p_{c,s,r,t} = \frac{\mu_{c,s,r,t}}{\sum_{s' \in \mathcal{S}_c} \mu_{c,s',r,t} + \varepsilon}, \qquad \varepsilon > 0, \tag{14}$$

and we write $\mathbf{p}_{c,r,t} = \{p_{c,s,r,t}\}_{s \in \mathcal{S}_c}$.

Diversity (normalized entropy). We quantify the dispersion of semantic composition by normalized entropy:

$$H_{c,r,t} = \frac{-\sum_{s \in \mathcal{S}_c} p_{c,s,r,t} \log p_{c,s,r,t}}{\log |\mathcal{S}_c|}. \tag{15}$$

B.4 Peak Year, Turning Point, and Reorganization

Peak year. The peak year is defined as the time slice maximizing concept magnitude:

$$t_{c,r}^{\mathrm{peak}} = \arg \max_{t \in \{1,\ldots,T\}} A_{c,r,t}. \tag{16}$$

Turning point and signed intensity. We define the turning point as the time slice with the strongest adjacent-slice change in magnitude:

$$t_{c,r}^{\mathrm{turn}} = \arg \max_{t \in \{2,\ldots,T\}} \left| A_{c,r,t} - A_{c,r,t-1} \right|, \tag{17}$$

and its signed intensity as

$$I_{c,r} = A_{c,r,t_{c,r}^{\mathrm{turn}}} - A_{c,r,t_{c,r}^{\mathrm{turn}}-1}. \tag{18}$$

Structural reorganization of composition. We quantify adjacent-slice reorganization of orientation shares by the L_1 change:

$$\begin{aligned} \Delta_{c,r,t} &= \left\| \mathbf{p}_{c,r,t} - \mathbf{p}_{c,r,t-1} \right\|_1 \\ &= \sum_{s \in \mathcal{S}_c} \left| p_{c,s,r,t} - p_{c,s,r,t-1} \right|. \end{aligned} \tag{19}$$

B.5 Implicit Concept Computation

Implicit vs. anchored subsets. Let $\mathcal{I}_{c,r}^{\mathrm{Imp}}$ denote salient implicit contexts and $\mathcal{I}_{c,r}^{\mathrm{Anch}}$ salient lexically anchored contexts, where the partition is determined by whether canonical lexemes of concept c are present.

Implicit-realization ratio. The implicit-realization ratio is

$$\bar{r}_{c,r} = \frac{\sum_{i \in \mathcal{I}_{c,r}^{\mathrm{Imp}}} m_i^{(c)}}{\sum_{i \in \mathcal{I}_{c,r}} m_i^{(c)}}. \tag{20}$$

B.6 Selecting Drifting Base Vectors

Time-slice mean activation. For a base vector k (SAE dimension), define its time-slice mean activation within a specified conditioning set $\mathcal{J}$:

$$\mu_{k,t} = \mathbb{E}[\, z_{i,k} \mid i \in \mathcal{J},\ t_i = t\,]. \tag{21}$$

Cumulative drift. We define cumulative drift as

$$D_k = \sum_{t=2}^{T} |\mu_{k,t} - \mu_{k,t-1}|\,. \tag{22}$$

We select base vectors with the largest D_k for interpretation. For diachronic evidence, each selected base is localized to the adjacent year pair $[y_1, y_2]$ where $|\mu_{k,t} - \mu_{k,t-1}|$ is maximal, and we retrieve the top-5 highest-activating sentences from y_1 and the top-5 highest-activating sentences from y_2.

B.7 Cross-Corpus Component Overlap

Corpus-conditioned drifting bases. For cross-corpus comparison, we use the salient set as the conditioning set, i.e., $\mathcal{J} = \mathcal{I}_{c,r}$, and compute $D_k^{(c,r)}$ by applying Eq. equation 21–equation 22 within each corpus r.

Top-K drifting set and Jaccard@K. Let $\mathcal{T}_{c,r}^K$ denote the set of the K base vectors with the largest $D_k^{(c,r)}$, with $K = 30$ in all reported experiments. For two corpora r and r', we measure overlap as

$$\mathrm{Jaccard@K}(c; r, r') = \frac{|\mathcal{T}_{c,r}^K \cap \mathcal{T}_{c,r'}^K|}{|\mathcal{T}_{c,r}^K \cup \mathcal{T}_{c,r'}^K|}. \tag{23}$$

This induces a decomposition into shared components $\mathcal{T}_{c,r}^K \cap \mathcal{T}_{c,r'}^K$ and corpus-specific components via set differences.

B.8 Cross-Layer Robustness Metrics

Layer-conditioned reruns. Let $\ell \in \mathcal{L}$ index SAE layers used in robustness tests (e.g., $\mathcal{L} = \{06, 14, 22, 29\}$). For each ℓ, we rerun the full pipeline with only the SAE layer switched, producing layer-specific trajectories $A_{c,r,t}^{(\ell)}$ and $\mathbf{p}_{c,r,t}^{(\ell)}$.

Peak and turning years across layers. Layer-specific peak and turning years are computed by applying Eq. equation 16 and Eq. equation 17 (and intensity by Eq. equation 18) to $A_{c,r,t}^{(\ell)}$.

Evidence-context 2-gram fingerprint and Avg. Jaccard. Let $\mathcal{E}_{c,r}^{(\ell)}$ be the collection of evidence contexts retrieved for interpretation under layer ℓ, using for each selected drifting base the peak adjacent year pair and the top-5 highest-activating sentences from each of the two years. Define the layer-specific fingerprint as the set of character 2-grams extracted from $\mathcal{E}_{c,r}^{(\ell)}$:

$$\mathcal{F}_{c,r}^{(\ell)} = \mathrm{2GramSet}\left(\mathcal{E}_{c,r}^{(\ell)}\right). \tag{24}$$

For two layers ℓ and ℓ', the 2-gram Jaccard similarity is

$$J\left(\ell, \ell'; c, r\right) = \frac{|\mathcal{F}_{c,r}^{(\ell)} \cap \mathcal{F}_{c,r}^{(\ell')}|}{|\mathcal{F}_{c,r}^{(\ell)} \cup \mathcal{F}_{c,r}^{(\ell')}|}. \tag{25}$$

We report the average cross-layer agreement for layer ℓ as

$$\mathrm{AvgJaccard}(\ell; c, r) = \frac{1}{|\mathcal{L}| - 1} \sum_{\substack{\ell' \in \mathcal{L} \\ \ell' \neq \ell}} J\left(\ell, \ell'; c, r\right). \tag{26}$$

B.9 Humanistic Interpretation Protocol

Methodological stance. We treat computational signals as methodological scaffolding rather than self-sufficient explanations.

Context retrieval. For selected orientations and drifting base vectors, we retrieve highest-activating evidence contexts with task-specific rules. For diachronic interpretation, we use each selected base's peak adjacent year pair and retrieve the top-5 highest-activating sentences from each year. For cross-corpus interpretation, we use the full-range top-30 highest-activating sentences for each base as the evidence pool and display 8 representative sentences per base. Concretely, we use the retrieved evidence contexts $\mathcal{E}_{c,r}$ as the primary material for interpretation.

Expert annotation and discursive roles. Domain experts assign semantic descriptions and articulate discursive roles grounded in the retrieved contexts. To prioritize base vectors most informative for diachronic reading, we select those with maximal cumulative drift (Eq. equation 22) over the period of interest and interpret them through expert analysis of their activating texts.

C Historian Validation of Implicit Concept Computation

Validation setup. To test whether anchor-absent high-activation cases genuinely instantiate the target concepts, we use a historian-annotated validation set derived from sampled implicit cases in the SAE activation space. Each concept contributes exactly 20 annotated cases in which the canonical concept word is absent. For each case, the historian assigns a binary judgment indicating whether the sentence treats the target concept as an implicit semantic support, underlying conceptual frame, or cognitive inertia. Table 3 reports the concept-level semantic-consistency rates. The detailed examples below focus on one representative base for each concept and, where relevant, note how many of the 20 concept-level cases fall under that base.

Concept	Cases	Consistency	Selected base	Base rate
society	20	90%	Base 96661 (*Party Linkage and Organizational Alignment*)	100%
nation	20	100%	Base 202680 (*Nation-State as Strategic Instrument*)	100%
world	20	100%	Base 81379 (*Revolutionary International Field*)	100%
individual	20	75%	Base 173164 (*Individualism as Discourse*)	100%

Table 3: Historian validation summary for sampled implicit cases. Each concept is evaluated on 20 annotated anchor-absent sentences. Concept-level rates are computed from the binary judgments in the annotated validation set; the final column reports the within-base rate for the representative base selected for detailed illustration below.

C.1 Selected Detailed Cases

***Society*: Base 96661 (*Party Linkage and Organizational Alignment*).** This base captures implicit uses of *society* as a field of class alignment, organized political coordination, and revolutionary transformation. Within the 20-case concept-level sample for *society*, five cases came from this base, and all five were judged semantically consistent.

1. ***The Guide*, 1926-01-21** `act`=10.7974

   东西各国的共产党和共产国际，应当联合团结一切劳动平民的革命力量和被压迫民族，一致反抗帝国主义而推翻他，推翻世界各国的资本主义，因为如果不是这样，不但无产阶级不能得著解放，就是弱小民族也始终不能脱离压迫。

   Translation. The communist parties of East and West, together with the Communist International, should unite all revolutionary forces among laboring people and oppressed nations, jointly resist imperialism, and overthrow the capitalism of all countries; otherwise, not only will the proletariat fail to gain liberation, but weak nations will never escape oppression.

   Analysis. The sentence construes social reality as an organized coalition of laboring people and oppressed nations united against imperialism and capitalism. Although the lexical form *society* does not appear, the passage presupposes a modern social totality structured by class relations, collective mobilization, and transnational alignment. In that sense, *society* functions as the underlying conceptual frame of the argument.

2. ***The Guide*, 1922-12-23** `act`=10.7117

   战后，保加利亚资产阶级政权解纽，共产党运动遂异常得势。

   Translation. After the war, the Bulgarian bourgeois regime disintegrated, and the communist movement accordingly gained extraordinary strength.

   Analysis. Here political change is explained through the collapse of a bourgeois regime and the corresponding rise of communist forces. The sentence therefore treats social order as a field whose structure is determined by class antagonism and can be reorganized through revolutionary struggle. This is a clear implicit realization of *society* as a historically transformable social formation.

***Nation*: Base 202680 (*Nation-State as Strategic Instrument*).** This base captures implicit uses of *nation* through arguments about state interest, political alignment, and the legitimacy of revolutionary versus counterrevolutionary power. In this validation set, all 20 anchor-absent cases sampled for *nation* came from this base, and all 20 were judged semantically consistent.

1. ***New Youth*, 1925-06-01** `act`=18.9140

   因此，我们研究了统治阶级的世界形势，还要研究被统治阶级的世界形势。换言之，认识了反革命势力，还须认识革命势力。

   Translation. Therefore, we have studied the world situation of the ruling classes, and we must also study the world situation of the ruled classes. In other words, having recognized the counterrevolutionary forces, we must also recognize the revolutionary forces.

   Analysis. Although the lexical form *nation* is absent, the passage decomposes political order into opposed forces whose legitimacy depends on the interests they represent. It implicitly treats the modern state as a political instrument embedded in class struggle rather than a morally unified whole. This makes statehood, and thus

the modern nation-state frame, the underlying structure of the argument.

2. ***The Guide*, 1926-01-07** `act`=18.5290

这些聪明人不懂得：（一）他们的劝告乃是完全取消列宁主义，因为这种政策等于放弃了世界革命的策略；（二）在「西方」帝国主义者（即他们劝告我们与之联合的）的观点看来，我们（苏联）即是「东方」不过比中国加倍「危险」罢了；（三）在对于东方几万万人大运动的关系之问题中，「中立」是不可能的，不管我们愿意中立或不愿意。

Translation. These clever people do not understand: first, their advice would amount to the complete abandonment of Leninism, because such a policy would mean giving up the strategy of world revolution; second, from the standpoint of the "Western" imperialists with whom they urge us to ally, we, the Soviet Union, are precisely the "East" only more dangerous than China; and third, with respect to the great movement of hundreds of millions in the East, neutrality is impossible, whether we desire it or not.

Analysis. This passage treats political actors as occupying irreducibly opposed positions within an international order defined by sovereignty, strategy, and alignment. It assumes that a state cannot stand outside the struggle between revolutionary and imperial forces, and that political legitimacy is inseparable from the interests a state chooses to embody. In this way, the sentence implicitly mobilizes a modern nation-state conception without naming it directly.

***World*: Base 81379 (*Revolutionary International Field*).** This base captures implicit uses of *world* as an international arena structured by competition, positionality, and large-scale political linkage. Within the 20-case concept-level sample for *world*, ten cases came from this base, and all ten were judged semantically consistent.

1. ***New Youth*, 1917-04-01** `act`=3.9910

吾国外交、素多弱点。欧战后国际地位、尤为可危。

Translation. Our country's diplomacy has long had many weaknesses. After the European war, its international position has become especially precarious.

Analysis. The passage interprets the European war as a global event and evaluates China in terms of its position within an international configuration. It therefore presupposes a modern world system composed of interdependent and dynamically shifting positions. Here, *world* operates as an implicit cognitive frame for diplomatic reasoning.

2. ***New Youth*, 1926-07-25** `act`=3.8709

美俄两条路的倾向，其基础亦在于美国和苏联对于国际政治的作用增高。

Translation. The tendency toward the American and Russian paths likewise rests on the heightened roles of the United States and the Soviet Union in international politics.

Analysis. The sentence explains ideological path choice through changes in the relative roles of the United States and the Soviet Union in international politics. It presupposes a world structured by great-power competition and differentiated political trajectories rather than a merely local field of debate. This makes *world* the implicit horizon within which the argument becomes intelligible.

***Individual*: Base 173164 (*Individualism as Discourse*).** This base captures implicit uses of *individual* through the public/private distinction, personal interest, and the attribution of political responsibility to discrete actors. Within the 20-case concept-level sample for *individual*, ten cases came from this base, and all ten were judged semantically consistent.

1. ***The Guide*, 1926-04-03** `act`=5.2608

方本仁对于江西民众的剥削搜括，也同别的军阀一样，分不出什么高低；他的失败，也同一切军阀的失败一样，私人军队，惟利是图，从前他在蔡承勋的部下推倒蔡承勋，现在又轮著他的部下邓如琢来推倒他了。

Translation. Fang Benren's extortion and exploitation of the people of Jiangxi were no better or worse than those of other warlords; his failure, too, was the same as that of all warlords: private armies, driven solely by profit. Once, while serving under Cai Chengxun, he overthrew Cai Chengxun; now it is his own subordinate Deng Ruzhuo's turn to overthrow him.

Analysis. The passage explains political failure through private armies, self-interest, and personal betrayal, drawing a sharp distinction between private motives and public order. It presupposes individuals as autonomous bearers of interest and responsibility rather than as mere nodes in a hierarchical moral order. On that basis, *individual* serves as the implicit explanatory framework for political criticism.

2. ***The Guide*, 1925-04-12** `act`=4.8186

蒋校长又时常向学生说：「军阀之所以成为军阀，全由于使其部下只知崇奉其私人而不知国家与人民；假使你们不服从党而服从蒋某一人，即是使我成为军阀，而终究要叛党叛国了」。

Translation. Principal Chiang also often told the students: "The reason warlords become warlords is precisely that they make their subordinates revere only their private person and not the nation and the people; if you obey not the Party but Chiang as an individual, you will turn me into a warlord, and in the end you will betray both the Party and the nation."

Analysis. This example turns on a sharp distinction between personal loyalty and public allegiance. Political legitimacy is assessed in terms of whether one is attached to a private individual or to public institutions such as the Party, the nation, and the people. Even without naming *individual* explicitly, the argument relies on a modern conception of the person defined by the public/private boundary and autonomous political responsibility.

C.2 Validation Against Prior Historiography

Setup. A stronger external validation than expert interpretation alone is to compare the turning years identified by the model against turning years recognized in prior historiography. Because historical scholarship more often periodizes these journals through corpus-level intellectual and political phases than through concept-specific quantitative dates, we use literature-backed corpus anchors as qualitative ground truth and test whether the strongest SAE turns align with them. For *New Youth*, we take 1919 as the standard historiographic turning point associated with the May Fourth break in the New Culture trajectory (Uberoi, 1972; Ni, 2021); for *The Guide*, we take 1923 as the turning phase associated with the consolidation of the Nationalist Revolution and the First United Front (Wilbur, 1984; Chen and Wei, 2005), and 1926 as the turning phase associated with the Northern Expedition and the reconfiguration of revolutionary politics (Jordan, 2019). The row-level *Source* column below records the specific historiographic support used for each corpus–concept pair.

Interpretation. Across all eight concept–corpus pairs, the strongest turning year identified by the SAE falls within one year of the historiography-backed reference year. This result suggests that the turning-point signal is not merely an internal property of the model's activation geometry, but also tracks historically recognizable phase changes. Where the alignment is off by one year, a plausible explanation is that changes in discursive structure need not coincide exactly with the public outbreak or retrospective naming of a major historical event, and may emerge slightly earlier or later in periodical discourse.

Limitations and future work. This validation remains small-scale and uses corpus-level historiographic anchors rather than concept-specific qualitative annotations, because curated historical ground truth is labor-intensive to assemble. A natural next step is to build a broader benchmark resource for historiography and historical newspaper research, pairing explicit qualitative judgments with reproducible temporal anchors. Such a resource could itself become useful infrastructure for AI for history.

C.3 Supplementary Comparison with DTM-lite (online LDA) and Static Word Vectors

Setup. We additionally compare the SAE-based pipeline against two supplementary baselines on the same sentence-level *New Youth* corpus (1915–1926) for the anchor concept *individual* (个人). This comparison focuses on two properties that can be summarized compactly across methods: implicit evidence capture and robustness under reruns. Here, *Implicit-evidence ratio* denotes the proportion of retrieved evidence cases that count as implicit evidence rather than lexically anchored evidence; *Evidence Avg. Jaccard* measures the agreement of retrieved evidence sets under reruns, using the same evidence-context fingerprint idea formalized in Appendix B.8; and *Turn robustness* is the agreement rate of the reported turning year under the relevant perturbation axis.

DTM-lite (online LDA). This baseline is a sentence-level bag-of-words topic-model pipeline, motivated by dynamic topic modeling (Blei and Lafferty, 2006) but implemented as online LDA (Hoffman et al., 2010). We tokenize with jieba, remove a fixed stopword list, build a corpus-wide count vocabulary with `min_df`=5 and `max_df`=0.8, and update an online LDA model year by year with $K = 40$ topics over seeds $0 \ldots 9$. For year t, we choose the topic with the largest $p(\text{anchor} \mid \text{topic})$ for anchor 个人.

Corpus	Concept	Strongest Turn (year, I)	Reference Turning Year	Source	$\lvert\Delta\rvert$	Acc@$\pm$1
New Youth	*individual*	1918 ($+0.2260$)	1919	Uberoi, 1972; Ni, 2021	1	1
New Youth	*nation*	1918 ($+0.1160$)	1919	Uberoi, 1972; Ni, 2021	1	1
New Youth	*society*	1918 (-0.2130)	1919	Uberoi, 1972; Ni, 2021	1	1
New Youth	*world*	1918 ($+0.2300$)	1919	Uberoi, 1972; Ni, 2021	1	1
The Guide	*individual*	1923 ($+0.0665$)	1923	Wilbur, 1984; Chen and Wei, 2005	0	1
The Guide	*nation*	1923 (-0.0810)	1923	Wilbur, 1984; Chen and Wei, 2005	0	1
The Guide	*society*	1924 ($+0.1520$)	1923	Wilbur, 1984; Chen and Wei, 2005	1	1
The Guide	*world*	1926 (-0.0852)	1926	Jordan, 2019	0	1

Table 4: Validation of strongest turning years against corpus-phase turning points recognized in prior historiography.

Method	Implicit-evidence ratio	Evidence Avg. Jaccard	Turn robustness
SAE (ours)	0.920	0.33–0.50 (across layers)	100%
DTM-lite (online LDA)	0.963	0.140	40%
Static word vectors (PPMI + SVD + Procrustes)	0.092	0.872	100%

Table 5: Supplementary comparison on *New Youth* for the anchor concept *individual* (个人). SAE reruns vary Layers 06/14/22/29 and use top-20 evidence sentences per run; DTM-lite (online LDA) and static word vectors (PPMI + SVD + Procrustes) rerun 10 random seeds and also use top-20 evidence sentences per run.

Sentence-level concept strength is the document-topic weight θ_{d,k_t} on that topic; yearly strength is the mean of this score over all / explicit / implicit sentences, where the explicit–implicit split is determined by whether the sentence contains the anchor substring. Evidence sentences are the top-20 sentences ranked by this score.

Static word vectors (PPMI + SVD + Procrustes). This baseline uses count-based diachronic word vectors (Levy et al., 2015) with cross-year alignment in the spirit of prior diachronic embedding work (Hamilton et al., 2016). For each year, we build a window-size-5 co-occurrence matrix, convert it to PPMI, factorize it with TruncatedSVD ($d = 300$), and align yearly spaces to a reference year using orthogonal Procrustes over the top-500 shared frequent words. Runs use seeds $0 \ldots 9$, `min_count`=5, and a vocabulary cap of 50,000 words. The concept vector for 个人 is the aligned seed-word vector; sentence vectors are obtained by mean pooling; sentence-level concept strength is cosine similarity to the year-specific concept vector; and yearly strength is the mean score over all / explicit / implicit sentences. Main results use reference year 1915, with 1920 as a sensitivity run. Evidence sentences are the top-20 by concept score, and drift evidence is the top-20 by drift score. This pipeline additionally applies a text-quality filter during evidence selection.

Interpretation. DTM-lite attains a high implicit-evidence ratio, but its retrieved evidence is markedly less stable and its reported turning point is sensitive to random initialization. Since this baseline is an online-LDA sentence bag-of-words approximation rather than the original Dynamic Topic Model of Blei and Lafferty (2006), its results are best read as a lightweight topic-model comparison. Static word vectors (PPMI + SVD + Procrustes) yield highly stable evidence sets, but their top-scoring cases are overwhelmingly driven by explicit lexical anchoring, so implicit capture is weak. Within this supplementary comparison, SAE occupies the more balanced regime: it preserves strong implicit capture while maintaining fully stable turning points and a reproducible evidence chain across layer or seed reruns.

Scope. These baselines have different inductive biases and retrieval details, and the static word-vector pipeline additionally applies a text-quality filter during evidence selection. We therefore treat this table as a supplementary robustness probe rather than a perfectly matched head-to-head benchmark. Other advantages of the SAE-based framework, especially stable concept decomposition and unified cross-concept / cross-corpus comparison, are better demonstrated through the interpretive analyses reported in the main text.

D Semantic Label Index and Base-Vector Evidence

To reduce cognitive load in Section 4, the main text refers to recurrent SAE components by short semantic labels rather than by raw base IDs. Table 6 maps each label to its constituent base(s). The compact entries below then provide (i) a semantic orientation, (ii) diachronic metrics when a stable single-base summary is available, and (iii) representative bilingual evidence. For cluster labels used in the multi-concept comparison, the ev-

idence is illustrative of the shared semantic pole captured by the constituent bases rather than an exhaustive profile of every base in the cluster.

D.1 Individual

***Actorhood* (Base 90370). Semantic orientation.** This label foregrounds the individual as an agentive subject marked by independence, initiative, self-respect, and responsibility.

Diachronic metrics. *New Youth*: `cum_drift` = 10.390, `peak_delta` = 2.490, `peak_years` = [1923, 1924]. *The Guide*: `cum_drift` = 2.705, `peak_delta` = 1.604, `peak_years` = [1922, 1923].

Representative evidence. New Youth

1. **`act`=23.6597**
   一曰损坏个人独立自尊之人格一。曰窒碍个人意思之自由。
   First, it damages the character of the individual as independent and self-respecting. Second, it obstructs the freedom of individual volition.
2. **`act`=23.1132**
   因此我们可以得到结论：（1）创作不宜完全没煞自己去模仿别人；（2）个性的表现是自然的并非由于民族主义等的主张（3）个性是个人唯一的所有，而又与人愿有根本上的共通点；（4）个性就是在可以保存范围内的国粹，有个性的新文学便是这国民所有的真的国粹的文学。
   From this we may draw the following conclusions: first, creative work should not efface the self completely in order to imitate others; second, the expression of individuality is natural and does not arise from doctrines such as nationalism; third, individuality is the only thing that belongs uniquely to the person, though it also shares fundamental common points with others; and fourth, individuality is the national essence insofar as that essence can be preserved, and a new literature with individuality is the genuine literature of national essence possessed by this people.
3. **`act`=22.6586**
   让个人自由发展，可以鼓励个人冒险、竞争、奋斗的精神，可以减少懒惰、不进取的脾气。
   Allowing the individual to develop freely can encourage the spirit of adventure, competition, and struggle, and can reduce habits of laziness and lack of initiative.

The Guide

1. **`act`=22.1041**
   很简单的理由，就是一些野心的军官，为了自己的利益，尽可以暂时依附革命旗下，但借此达到了个人目的以后，还会管革命事业吗?
   The reason is very simple: some ambitious officers may temporarily attach themselves to the revolutionary banner for their own interests, but once they have achieved their personal ends in this way, will they still care about the revolutionary cause?
2. **`act`=19.9043**
   现在国民党用来革命的军队，多半是随时募集或改编收容的，不特兵士不知道革命的意义；就是军队的领袖除了个人活动的欲望之外，多半不了解或服从主义。
   The armies that the Nationalist Party now uses for revolution have mostly been hastily recruited or reorganized and absorbed. Not only do the soldiers fail to understand the meaning of revolution, but even among the military leaders, apart from the desire for personal advancement, most neither understand nor obey the doctrine.
3. **`act`=20.7651**
   你们竟想以伪和平的假面具，掩饰你们个人权利禄位的贪心！
   You would actually use the false mask of peace to conceal your greed for personal power, privilege, and office!

***Individualism as Discourse* (Base 173164). Semantic orientation.** This label captures an explicit discourse of individualism, especially arguments about personal autonomy, individual freedom, and the proper limits of political intervention.

Diachronic metrics. *New Youth*: `cum_drift` = 15.747, `peak_delta` = 5.078, `peak_years` = [1924, 1925]. *The Guide*: `cum_drift` = 2.596, `peak_delta` = 1.131, `peak_years` = [1923, 1924].

Representative evidence. New Youth

1. **`act`=16.3406**
   下次讲个人主义—自由主义—一派的坏处，在于把国家的势力太限制了，以为国家只可维持关于物质方面的平安，他的权力，愈小愈好。
   Next time, when discussing the faults of the school of individualism-liberalism, its defect lies in restricting the power of the state too severely, taking the view that the state should do no more than maintain material peace and that the smaller its power, the better.
2. **`act`=15.5903**

Main-text label	Constituent base(s)	Concept	One-line gloss
Actorhood	90370	*individual*	The individual as an agentive subject of initiative, self-direction, responsibility, and public action.
Individualism as Discourse	173164	*individual*	Individualism as an explicit discursive register of autonomy, freedom, and the limits of state coercion.
Property and Economic Individuality	206475	*individual*	The individual in relation to contract, production, property, and economic coordination.
Societal Transition and Institutional Design	3810	*society*	Society as a site of structural reorganization, institutional redesign, and transitional ordering.
Organized Praxis and Labor-Movement Alignment	25413 + 224715	*society*	Society articulated through labor-movement line struggle, socialist organization, and practical mobilization.
Party Linkage and Organizational Alignment	104017 + 96661 + 63228	*society*	Society articulated through party coordination, class alignment, and communist organizational linkage.
Nation-State as Strategic Instrument	202680	*nation*	The nation-state treated instrumentally through state form, class basis, and geopolitical alignment.
Revolutionary International Field	81379	*world*	The world as a structured international arena of revolution, competition, and positionality.

Table 6: Semantic labels used in the main text and their mapping to the underlying SAE base(s).

而且个人或小团体绝对自由，则生产额可以随意增减，有时社会需要多而生产少，有时需要少而生产多，因为没有统一机关用强制力去干涉调节，自然会发生生产过剩或不足的弊端。

Moreover, if individuals or small groups enjoy absolute freedom, the amount produced may increase or decrease at will: at times society needs more while less is produced, and at times it needs less while more is produced. Because there is no unified organ using coercive power to intervene and regulate, the evils of overproduction or insufficiency naturally arise.

3. **act=14.8172**
个人主义的中心观念，便是根据个人自由意志商定契约，不要政府用法律的或政治的势力去干涉他们，只听他们自由去做。
The central idea of individualism is to conclude contracts according to individual free will and not allow the government to interfere through legal or political power, but simply let people act freely.

The Guide

1. **act=14.0948**
我们应该竭诚忠告晨报记者，个人立言错了是小事，因为要回护自己的错遂不顾社会的错是大事；因为不忍社会的错遂不惜承认自己的错，这是最勇敢的行为呵！
We should sincerely advise the Morning Post reporter that it is a small matter for an individual to speak wrongly, but it is a grave matter to disregard society's error merely in order to defend one's own. To acknowledge one's own error because one cannot bear society's error, by contrast, is the most courageous conduct.

2. **act=12.2118**
据弟个人观察，全出于如下误会：即以为阶级争斗，即是劳工专政，劳工专政，即是想将劳工阶级一变为压迫人的阶级。
According to my personal observation, it all arises from the following misunderstanding: that class struggle is taken to mean the dictatorship of labor, and the dictatorship of labor is taken to mean transforming the laboring class into a class that oppresses others.

3. **act=10.8216**
这不是我信口瞎说，都有事实可以证明的。我个人就是一个例。
This is not something I say at random; there are facts to prove it. I myself am one example.

***Property and Economic Individuality* (Base 206475). Semantic orientation.** This label isolates the economic pole of the *individual* concept: contract, property, production, and the problem of whether individual discretion can organize economic life. A recurrent theme in the retrieved evidence is that individually owned modes of production cease to satisfy the needs of social development.

Diachronic metrics. *New Youth*: `cum_drift` $= 4.496$, `peak_delta` $= 1.125$, `peak_years` $= [1923, 1924]$.

Representative evidence. New Youth

1. **act=15.5903**
而且个人或小团体绝对自由，则生产额可以随意增减，有时社会需要多而生产少，有时需要少而生产多，因为没有统一机关用强制力去干涉调节，自然会发生生产过剩或不足的弊端。
Moreover, if individuals or small groups enjoy absolute freedom, the amount produced may increase or decrease at will: at times society needs more while less is

produced, and at times it needs less while more is produced. Because there is no unified organ using coercive power to intervene and regulate, the evils of overproduction or insufficiency naturally arise.

D.2 Society

***Societal Transition and Institutional Design* (Base 3810). Semantic orientation.** This label marks the transition-oriented pole of *society*: social reorganization, institutional redesign, and the claim that revolutionary or developmental change requires restructuring social order rather than merely renaming it.

Representative evidence. New Youth

1. **1922-07-01; `act`=11.4612**
   （2）关于政治教育，社会主义的青年应宣传社会主义于大多数青年无产阶级，其方法或集会讲演，或刊行出版物和小册子，并特别讲述中国政治情形及其他种种情形，以启发并养成青年无产阶级的政治觉悟及批评力。
   (2) On political education: socialist youth should propagate socialism among the broad masses of young proletarians, whether through meetings and lectures or through the publication of periodicals, pamphlets, and booklets, and should in particular explain China's political conditions and various other conditions so as to awaken and cultivate the political consciousness and critical capacity of the young proletariat.
2. **1921-04-01; `act`=11.0933**
   这婚姻法在法律上实现男女的绝对平等，由资本主义到社会主义的过渡期的状态中，给妇女以可能的范围内的自由，离婚则由男女双方合意或者单由一方的意思，亦可实行，父母对于子女的权利义务双方平等，因此打破旧结婚制度，同时作成未来男女关系更为自由的基础。
   This marriage law legally realizes absolute equality between men and women. In the transitional state from capitalism to socialism, it grants women freedom to the fullest extent possible; divorce may be carried out either by mutual consent or by the will of one party alone; and both father and mother possess equal rights and obligations with respect to their children. It thereby breaks the old marriage system while laying the foundation for freer relations between men and women in the future.
3. **1926-05-25; `act`=10.8985**
   因为农村中社会主义建设的根本道路，就在，于社会主义的国家工业、国家信托机关以及在无产(53)阶级手里其他机关之经济指导权的增长底下，引导农民根本群众进于协作社的组织，并保证这种组织之社会主义的发展，利用、制服并限制其资本主义的原素。
   For the fundamental path of socialist construction in the countryside lies precisely in, under the growth of the economic directing power of socialist state industry, state trusts, and other organs in the hands of the proletariat, guiding the broad masses of peasants into cooperative organization, guaranteeing the socialist development of such organization, and making use of, subduing, and restricting its capitalist elements.

The Guide

1. **1924-06-18; `act`=9.6285**
   第一，就国民党的主义上讲：此时任何政党党纲，都论列到社会的经济政策，可是中国国民党二十年前造端时即注意到民生问题，这是受了德法两国劳动运动的影响，而后进的中国国民党遂有此特色——和民族民权并列的民生主义。
   First, in terms of the doctrine of the Nationalist Party: at this moment every party program discusses the economic policy of society, yet when the Chinese Nationalist Party was founded twenty years ago it had already taken note of the livelihood question. This was due to the influence of the labor movements in Germany and France, and it is in this sense that the belated Chinese Nationalist Party acquired this characteristic: the Principle of People's Livelihood placed alongside nationalism and civil rights.
2. **1924-09-17; `act`=8.9077**
   劳资斗争是社会进化上一种不可免的革命现象，主张劳资调和是一种和缓革命的政策，无人能够相信不革命的调和政策可以平均地权，可以限制资本，世界上那有这样好说话的大地主与资本家?
   Labor-capital struggle is an unavoidable revolutionary phenomenon in the evolution of society. To advocate labor-capital harmony is a policy for softening revolution. No one can believe that a harmonizing policy that avoids revolution can equalize land rights or restrain capital; where in the world are there such accommodating landlords and capitalists?
3. **1923-04-18; `act`=8.1119**
   劳动者的组织一天天的发达，一天天的集中，在中国社会上已成为一种新势力。
   Laborers' organizations are developing and concentrating day by day; within Chinese society they have already become a new force.

***Organized Praxis and Labor-Movement Alignment* (Bases 25413 + 224715). Semantic orientation.** This cluster label captures society as it is articulated through labor-movement organization, socialist line struggle, and practical revolutionary alignment. Within this cluster, Base 224715 is the clearest evidence-bearing pole.

Diachronic metrics (documented constituent Base 224715). *New Youth*: `cum_drift` = 6.019, `peak_delta` = 2.162, `peak_years` = [1924, 1925]. *The Guide*: `cum_drift` = 5.020, `peak_delta` = 1.818, `peak_years` = [1923, 1924].

Representative evidence. New Youth

1. **`act=11.6260`**
   所以对于初期的社会主义，乌托邦的共产主义，不识时务穿著理思的绣花衣裳的无政府主义，专主经济行动的工团主义，调和劳资以延长资本政治的吉尔特社会主义，以及修正派的社会主义，一律排斥批评，不留余地。
   Therefore, with regard to early socialism, utopian communism, anarchism that is out of season and dressed in the embroidered garments of idealism, syndicalism that specializes in economic action, guild socialism that reconciles labor and capital so as to prolong capitalist politics, and revisionist socialism, we reject and criticize them all alike, leaving no room.
2. **`act=10.9037`**
   各国的改良主义者，第二国际的社会党叛徒和机会主义者，各国的孟雪维克党人虚伪地戴著马克思和恩格斯学理的假面具。
   The reformists of all countries, the social-traitors and opportunists of the Second International, and the Menshevik party members of every country hypocritically wear the false mask of the doctrines of Marx and Engels.
3. **`act=10.0740`**
   所以我们要反对伦敦会议，专家计划，固然须反对国际资本帝国主义，但亦须反对无产阶级底叛徒：社会党、劳动党、社会民主党——一句话，第二国际！
   Therefore, if we are to oppose the London Conference and the experts' plan, we must certainly oppose international capitalist imperialism, but we must also oppose the traitors to the proletariat: the Socialist Party, the Labour Party, the Social Democratic Party, in a word, the Second International!

The Guide

1. **`act=10.8509`**
   （二）左派大联合（Bloc des Gauches）这一派包含有左派共和，社会主义共和，急进，社会主义急进等左派政团，近年以来，社会党亦与之勾结，因为凡全国之主张较左者均集合于此，故亦称如左派全国大联合（Bloc National de gauche）他的主张虽比前派为急进，然不过带点改良色彩罢了。
   (2) The Left Bloc (Bloc des Gauches) includes left republicans, socialist republicans, radicals, socialist radicals, and other left-wing political groups. In recent years the Socialist Party has also aligned itself with it. Because all those in the country whose views are comparatively left-leaning gather here, it is also called the National Left Bloc (Bloc National de gauche). Although its program is more radical than that of the previous camp, it is after all only tinged with reformism.
2. **`act=10.1573`**
   工人阶级这种失败的原因，大半在于运动之中有第二国际派的社会党，他们受资产阶级的利用，破坏工人阶级斗争的阵线，几乎使工会的国际联合（职工国际）变成资本家的国际联合（国际联盟）的附庸。
   A large part of the reason for this failure of the working class lies in the presence within the movement of Socialist parties of the Second International. Used by the bourgeoisie, they break up the front of working-class struggle and almost turn the international union of trade unions into a vassal of the capitalist international union, the League of Nations.
3. **`act=9.0114`**
   但是当一九一四年七月底八月初，那可怖的消息传布之后，世界的大屠杀确已开始了后，第二国际名下的社会党竟翻过脸来，举起他们的赤帜，招呼党人投入敌人的营垒中替争权利的帝国主义者出死力了。
   But after the terrible news spread at the end of July and beginning of August 1914, when the world's great slaughter had indeed begun, the Socialist parties under the name of the Second International suddenly changed face, raised their red banners, and called upon their party members to throw themselves into the enemies' camp and sacrifice their lives for the imperialists contending over rights and interests.

***Party Linkage and Organizational Alignment* (Bases 104017 + 96661 + 63228). Semantic orientation.** This cluster emphasizes party coordination, class alignment, and communist organizational linkage. Within the historian-validated

implicit sample, the best-documented constituent base, 96661, reaches 100% semantic consistency; see Appendix C. The secondary constituent evidence from Base 63228 is used more sparingly, because it highlights factional and organizational alignment around nationalist and right-wing political blocs rather than the full semantic range of the cluster.

Representative evidence. New Youth

1. **1926-03-25; Base 63228; `act`=6.8142**
   固然——五卅之后，国民运动内部起了剧烈的阶级分化的现象，不但资产阶级直接的压迫束缚工人阶级而且政党界思想界，也因此而发生分化：这一分化开始于戴季陶先生的反对阶级斗争及所谓「右派国民党员站起来」的运动，结果是极右派利用戴季陶先生的领袖，召集西山会议——国民党中央委员会之右派会议；学生界里也发生所谓国家主义的运动，成立国家主义团体联合会。
   It is true that after the May Thirtieth Movement there arose within the national movement an intense phenomenon of class differentiation: not only did the bourgeoisie directly oppress and restrain the working class, but the worlds of parties and ideas also split accordingly. This differentiation began with Mr. Dai Jitao's opposition to class struggle and the movement for so-called right-wing Nationalist Party members to stand up; the result was that the extreme right, under Mr. Dai's leadership, convened the Xishan Conference, that is, the right-wing meeting of the Nationalist Party Central Executive Committee. In student circles there also emerged the so-called nationalist movement, and the United League of Nationalist Organizations was established.

The Guide

1. **1926-01-21; Base 96661; `act`=10.7974**
   东西各国的共产党和共产国际，应当联合团结一切劳动平民的革命力量和被压迫民族，一致反抗帝国主义而推翻他，推翻世界各国的资本主义，因为如果不是这样，不但无产阶级不能得著解放，就是弱小民族也始终不能脱离压迫。
   The communist parties of East and West, together with the Communist International, should unite all revolutionary forces among laboring people and oppressed nations, jointly resist imperialism, and overthrow the capitalism of all countries; otherwise, not only will the proletariat fail to gain liberation, but weak nations will never escape oppression.
2. **1922-12-23; Base 96661; `act`=10.7117**
   战后，保加利亚资产阶级政权解纽，共产党运动遂异常得势。
   After the war, the Bulgarian bourgeois regime disintegrated, and the communist movement accordingly gained extraordinary strength.
3. **1926-02-03; Base 63228; `act`=8.3208**
   国内许多政党和政派——如国民党右派的孙文主义学会，国家主义联合会，国家主义的醒狮周报，如今表面上也赞成国民会议。
   Many domestic parties and political groupings, such as the Sun Yat-senist Society of the right wing of the Nationalist Party, the Nationalist Federation, and the nationalist Awakened Lion Weekly*, now also superficially endorse the National Assembly.*

D.3 Nation and World

***Nation-State as Strategic Instrument* (Base 202680). Semantic orientation.** This label treats the modern nation-state instrumentally, foregrounding state form, class basis, and geopolitical alignment rather than an essentialized national spirit.

Diachronic metrics. *New Youth*: `cum_drift` = 6.786, `peak_delta` = 1.359, `peak_years` = [1923, 1924]. *The Guide*: `cum_drift` = 1.797, `peak_delta` = 1.045, `peak_years` = [1925, 1926].

Representative evidence. New Youth

1. **`act`=16.1241**
   （二）我以为世界上只有两个国家：一是资本家的国家，一是劳动者的国家，但是现在除俄罗斯外，劳动者的国家都还压在资本家的国家底下，所有的国家都是资本家的国家，我们似乎不必妄生分别。
   (2) I hold that there are only two kinds of states in the world: one is the state of the capitalists, and one is the state of the laborers. But at present, apart from Russia, the states of the laborers are still suppressed beneath the states of the capitalists; all states are states of the capitalists, and it seems that we need not draw idle distinctions.
2. **`act`=16.0973**
   第四；俄国底共产党和德国底社会民主党虽然同一不反对国家组织，是他们不同之点有三：（一）生产机关集中到国家手里，在共产党是最初的手段，在社会民主党是最终的目的；（二）德国社会民主党带著很浓的德意志国家主义的色采，俄国

共产党还未统一国内，便努力第三国际的运动；（三）社会民主党所依据的国家是有产阶级的国家，共产党所依据的国家是无产阶级的国家。

Fourth, although the Russian Communist Party and the German Social Democratic Party are alike in not opposing state organization, there are three points at which they differ: first, concentrating the organs of production in the hands of the state is for the Communists an initial means, but for the Social Democrats a final end; second, the German Social Democratic Party bears a heavy coloration of German statism, whereas the Russian Communists, before even unifying the country, already strove for the movement of the Third International; third, the state on which the Social Democrats rely is a state of the propertied classes, whereas the state on which the Communists rely is a state of the proletariat.

3. **`act=14.2716`**
   我承认国家只能做工具不能做主义，古代以奴隶为财产的市民国家，中世以农奴为财产的封建诸侯国家，近代以劳动者为财产的资本家国家，都是所有者的国家，这种国家底政治法律，都是掠夺底工具，但我承认这工具有改造进化的可能性，不必根本废弃他，因为所有者的国家固必然造成罪恶，而所有者以外的国却有成立的可能性；

   I admit that the state can only serve as an instrument and cannot serve as a doctrine. The civic states of antiquity, which treated slaves as property; the feudal states of the Middle Ages, which treated serfs as property; and the capitalist states of modern times, which treat laborers as property, are all states of owners. The politics and laws of such states are all instruments of plunder. Yet I admit that this instrument has the possibility of reform and evolution and need not be discarded root and branch, because while the state of owners necessarily produces evil, a state other than that of owners has the possibility of being established.

The Guide

1. **`act=15.4541`**
   现代所谓帝国主义乃指资本帝国主义，其存在须有下列二个特性：（一）凡是帝国主义的国家，无论大小强弱，必然是资本主义制度的国家；（二）凡是帝国主义的国家，其国内资本主义必然发展到财政资本主义向国外掠夺压迫殖民地及半殖民地。

   What is today called imperialism means capitalist imperialism, and its existence requires the following two characteristics: first, every imperialist state, regardless of size or strength, is necessarily a state with a capitalist system; second, within every imperialist state domestic capitalism must necessarily have developed into finance capital that plunders and oppresses colonies and semi-colonies abroad.

2. **`act=15.3258`**
   但先生忘记了：现在的中国还不是纯粹资产阶级统治下的独立国家，乃是外国帝国主义和封建余孽——军阀，统治下的半殖民地。

   But sir has forgotten this: present-day China is not yet an independent state under the rule of a pure bourgeoisie, but rather a semi-colony ruled by foreign imperialism and the remnants of feudalism, namely the warlords.

3. **`act=15.0802`**
   苏俄之所以赤，乃因为十月革命是工农阶级推翻资产阶级与资本主义的革命，一切资本帝国主义者正因此而仇视他；如果他现在也变成资本帝国主义的国家，那还何赤之有?

   The reason Soviet Russia is "red" is that the October Revolution was a revolution of workers and peasants overthrowing the bourgeoisie and capitalism, and all capitalist imperialists hate it for precisely this reason. If it were now also to become a capitalist-imperialist state, what redness would remain?

Revolutionary International Field **(Base 81379).** **Semantic orientation.** This label captures *world* as an international arena structured by revolution, competition, positionality, and large-scale geopolitical linkage. In the historian validation reported in Appendix C, the annotated implicit sample for this base reaches 100% semantic consistency.

Representative evidence. New Youth

1. **1926-07-25; `act=8.5350`**
   世界资本主义的稳定是相对的，还是绝对的呢?现在的资本主义发展与战前资本主义的发展有什么不同呢?这是决定世界革命运动消沉与否的前提。

   Is the stabilization of world capitalism relative or absolute? How does the current development of capitalism differ from its development before the war? This is the premise that determines whether the world revolutionary movement is in decline or not.

2. **1926-07-25; `act=6.7189`**
   中国的民族革命运动是世界革命运动的一部分，这句话的意义是说，中国民族革命

运动受世界革命运动的影响和辅助，并且它也可以影响世界革命运动。

China's national revolutionary movement is a part of the world revolutionary movement. The meaning of this statement is that China's national revolutionary movement is influenced and assisted by the world revolutionary movement, and that it can in turn influence the world revolutionary movement.

3. **1926-03-25; `act`=6.6014**
   再则，中国国民革命和世界的社会革命之联合战线，中国的民族解放运动和世界的无产阶级革命运动之联合战线也在这一次实现出来——苏联、英、法、德、日等无产阶级及其革命的政党，共产党，都奋起援助。

   Moreover, the united front between China's national revolution and the world's social revolution, and the united front between China's national liberation movement and the world's proletarian revolutionary movement, were also realized on this occasion: the proletarians and revolutionary parties, the communist parties, of the Soviet Union, Britain, France, Germany, Japan, and other places all rose to provide assistance.

The Guide

1. **1926-07-14; `act`=8.4024**
   中国民族解放运动，已成为世界革命的联合战线之一员，我们已经不是孤立的了！

   China's national liberation movement has already become a member of the united front of world revolution; we are no longer isolated!

2. **1925-03-21; `act`=7.3691**
   要中国革命成功，必须与世界革命运动即西方无产阶级的革命相联合，因为两者的敌人是共同的，两者的目的同是推翻资本帝国主义。

   For China's revolution to succeed, it must unite with the world revolutionary movement, that is, the revolution of the western proletariat, because the enemies of the two are common and the aim of both is the overthrow of capitalist imperialism.

3. **1926-05-15; `act`=6.5169**
   这句话的意义，是事实逼迫著他们不能不认识中国的民族解放运动是世界的而不是国家的了。

   The meaning of this statement is that the facts have forced them to recognize that China's national liberation movement is of the world rather than merely of the nation.