Serican Journal of Medicine
2026, VOL. 3, NO. 2, 25497
https://doi.org/10.17161/sjm.v3i2.25497

BERTopic Mapping of Clinical Teaching Activities in Residency and Medical Internship Training

Chujie Chen and Jun Li*

Department of Urology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 518000, China
Corresponding Author: Jun Li, MD/PhD, Department of Urology, The Seventh Affiliated Hospital, Sun Yat-sen University, 628 Zhenyuan Road, Guangming District, Shenzhen 518000, China. Email: lijun273@mail.sysu.edu.cn. ORCID: https://orcid.org/0000-0003-1227-6939

ABSTRACT

Background: Electronic teaching platforms record thousands of free-text descriptions of clinical teaching activities, but rarely at scale. We applied BERTopic to discover the topic structure of teaching narratives in two parallel training systems and subjected the taxonomy to a multi-LLM zero-shot encoding-reproducibility check using independent commercial large language models (LLMs).

Methods: We analyzed 4,811 de-identified records (2,890 residency, 1,921 medical internship) from the CCMTV platform at a single-center tertiary teaching hospital (2022-2025). After jieba tokenization and 46,313 PHI placeholder substitutions, BERTopic with BAAI/bge-base-zh-v1.5 embeddings was fit per system and reduced to 24 topics each. Coherence (c_v, NPMI), topic diversity, 5-seed bootstrap (ARI/NMI), and a min_cluster_size grid characterized robustness. Two LLMs (Codex GPT-5.2; Gemini 3) annotated 200 stratified-sampled records under identical zero-shot prompts; inter-LLM Cohen's κ quantified encoding reproducibility. Cross-system topic correspondence used cosine similarity of topic centroids with Hungarian matching.

Results: Reduced models reached c_v 0.505-0.548, diversity 0.86-0.90, outlier rates 14-23%, ARI 0.60-0.67, NMI ≥ 0.90. Inter-LLM Cohen's κ = 0.709 [95% CI 0.64-0.78]; 0.878 [0.82-0.92] excluding -1 residuals. Cross-system Hungarian matches yielded cosine 0.783 ± 0.132; 17 of 24 pairs (71%) reached ≥ 0.70, but a 1000-permutation merged-relabel null showed this count was not significantly above chance, supporting descriptive interpretation only. Per-topic binomial GLM (year predictor, BH-FDR over 24 topics per system) identified significant year trends in 17 of 24 residency and 15 of 24 internship topics.

Conclusions: BERTopic with multi-LLM inter-LLM encoding-reproducibility checking offers a scalable, descriptive framework for monitoring the topic structure of clinical teaching and screening for potential topic-share changes, though population-level overlap claims should be read cautiously given the permutation-null result.

ARTICLE HISTORY
Received: May 13, 2026
Revised: May 27, 2026
Accepted: May 27, 2026

KEYWORDS
clinical teaching, topic modeling, BERTopic, natural language processing

Introduction

Clinical teaching activities form the operational backbone of postgraduate medical training. In China, the national standardized residency training system, established in 2014, mandates structured teaching for all new medical graduates [1, 2], and parallel medical internship rotations expose senior medical students to ward-based bedside instruction. Electronic teaching management platforms, such as the Cloud-based Clinical Medical Teaching and Visualization (CCMTV) platform widely deployed in Chinese tertiary hospitals, document each session with free-text narratives describing topic, learning objectives, methods, and reflections. These narratives accumulate into corpora of thousands of records per institution per year, yet beyond program-level descriptive statistics they are seldom mined at scale for latent topic structure. Existing reports from individual institutions typically rely on manual content review of small samples or descriptive frequencies of pre-

specified categories, approaches that scale poorly and risk imposing investigator-defined ontologies onto data that may have richer latent structure.

Unsupervised topic modeling offers an alternative path. Latent Dirichlet Allocation (LDA) has been applied to medical-education-adjacent text [3], but its bag-of-words assumption can be limiting for short, jargon-dense Chinese clinical text, motivating the use of contextualized embedding-based topic models such as BERTopic [4], a neural topic-modeling pipeline that combines transformer embeddings, UMAP dimensionality reduction, HDBSCAN density clustering, and class-based TF-IDF for topic representation. BERTopic has shown promising performance on medical and biomedical corpora [5,6]. However, two methodological gaps remain. First, validation: established practice for unsupervised topic taxonomies relies on manual expert coding to assess interpretability, an expensive and slow process that is the principal bottleneck in scaling such studies. Recent work suggests that commercial large language models (LLMs) can match or exceed crowd-worker reliability for routine text annotation tasks [7,8], raising the possibility of fast, reproducible LLM-based topic validation, but this has not yet been systematically applied to topic models of clinical teaching content. Second, comparison: most BERTopic studies of medical text fit a single corpus, rather than asking whether the topic structure recovered from parallel training programs (residency vs internship; specialist vs generalist learners) is shared, distinct, or partially overlapping, a question with direct implications for curriculum coordination.

We address both gaps in a dual-system corpus of free-text teaching narratives from the CCMTV platform at a tertiary teaching hospital in Shenzhen, China. We pursue four research questions: (1) What latent topic structure does BERTopic recover from residency and medical internship teaching narratives, and is the model coherent and reproducible? (2) To what extent do two independent commercial LLMs agree on topic assignment under zero-shot prompting, and does this agreement support the topic taxonomy as semantically interpretable? (3) How similar are topics discovered independently in the two training systems, and which topics are shared versus system-specific? (4) Has the topic distribution drifted over four annual cross-sections (2022-2025), and what shifts in clinical teaching content does this reveal? We do not propose LLM-vs-human gold-standard agreement; we explicitly frame inter-LLM agreement as between-model encoding reproducibility, not as substitution for human expert validation.

Methods

Study Design and Data Source

This was a retrospective text-analytic study of teaching activity records extracted from the CCMTV electronic teaching management platform at the Seventh Affiliated Hospital, Sun Yat-sen University (Shenzhen, China), covering January 2022 through December 2025. Figure 1 provides a graphical overview of the full analysis pipeline; readers may use it as a navigational map alongside the Methods subsections that follow. Two parallel modules were analyzed: the residency training module (postgraduate residents in standardized training) and the medical internship module (final-year medical students on clinical rotations). Each CCMTV record contains structured fields (date, department, training program / level, session type, speaker username) and three free-text fields, for residency: session title (name), learning objectives (introduce), and process record (record); for medical internship: session title (name), learning objectives (introduction), and activities record (activities_record). The three free-text fields were concatenated (newline-separated). To remove operator-introduced redundancy where a later field repeated text from an earlier field (a common operator pattern when the trainee fills the same content into both name and introduction fields), we performed within-record longest-common-substring exact-match removal across the three fields. The concatenated text was then processed through the layered PHI pipeline (next subsection) and collapsed to identical-text uniqueness across records. From the 3,619 residency records and 3,216 internship records in the raw export, we excluded records with empty text (residency 0; internship 9), records with concatenated text < 50 Chinese characters after tokenization (residency 381; internship 551), and exact duplicate text (residency 348; internship 735; the internship duplicate-text bucket includes 464 records where the introduction field was identical to the name field). After exclusions the analytical sample comprised 4,811 records (residency 2,890; medical internship 1,921) spanning 27 clinical departments.

Text Preprocessing and De-identification

Free-text narratives were processed in Python 3.13. We performed Chinese tokenization with the

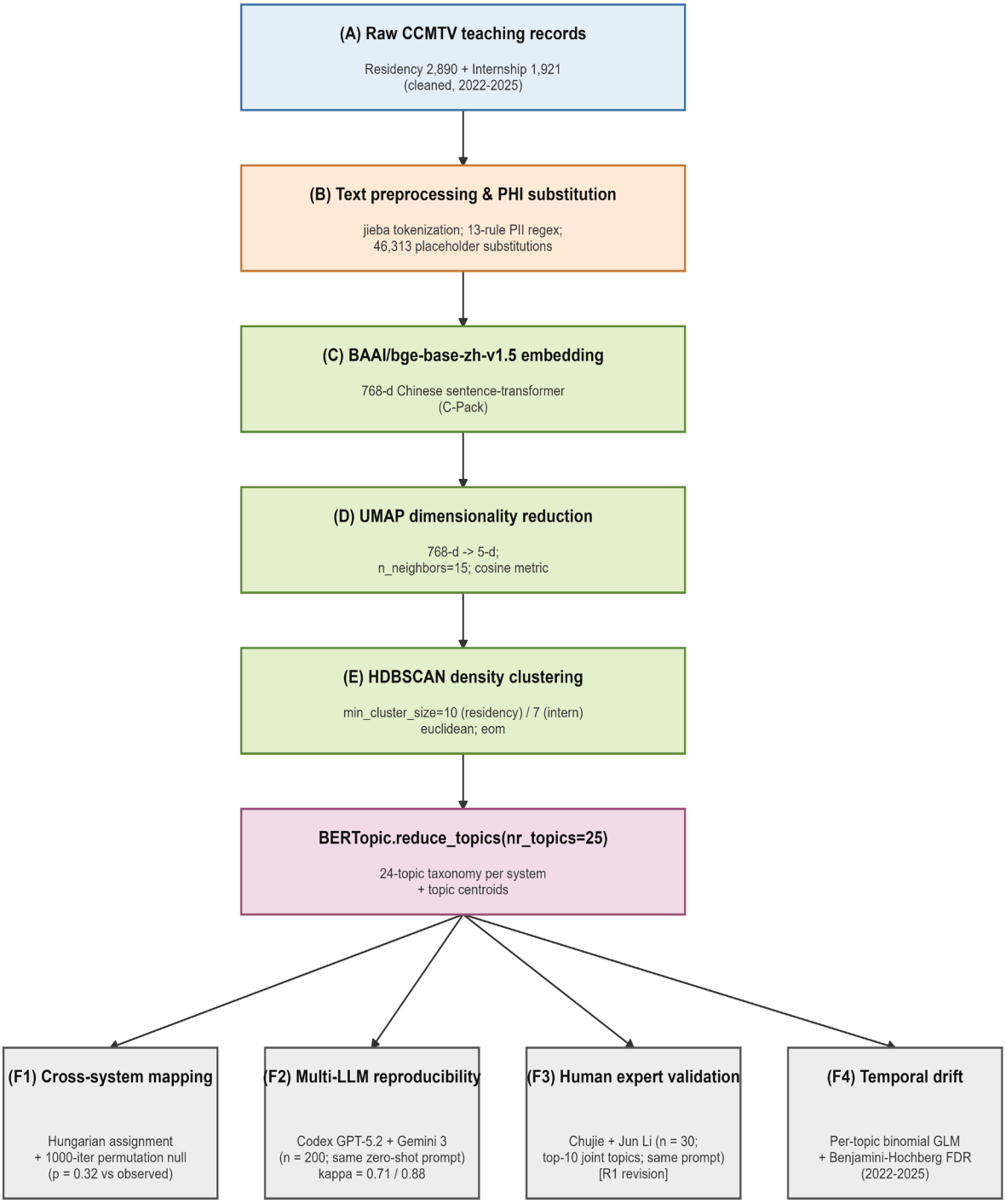

Figure 1. Analysis workflow diagram. Left-to-right reading order: (A) raw CCMTV teaching records (residency n = 2,890; internship n = 1,921 after cleaning) → (B) text preprocessing and PHI substitution (46,313 placeholder substitutions across the corpus; PII v3 13-rule layered regex; Supplementary Methods §1) → (C) BAAI/bge-base-zh-v1.5 sentence-transformer embedding (768-d) → (D) UMAP dimensionality reduction (n_components = 5, cosine metric) → (E) HDBSCAN density clustering (min_cluster_size = 10 residency / 7 internship) → BERTopic.reduce_topics(nr_topics = 25) → 24-topic taxonomy per system + topic centroids; followed by four downstream analyses: (F1) cross-system Hungarian matching with 1000-iteration merged-relabel permutation null (Methods §"Cross-System Topic Mapping"); (F2) multi-LLM zero-shot encoding reproducibility (Codex GPT-5.2 + Gemini 3, 200 stratified-sampled records, Cohen's κ); (F3) human expert validation sub-study on a 30-record subset (Methods §"Human Expert Validation"); and (F4) per-topic temporal drift via binomial GLM with Benjamini-Hochberg FDR over 2022-2025 annual cross-sections. Each box is annotated with the corresponding Methods subsection.

jieba 0.42 segmenter, augmented with a domain-specific medical and pedagogical lexicon of 50 terms (e.g., "腰椎间盘突出" [lumbar disc herniation], "教学查房" [teaching ward rounds]); the complete lexicon is reproduced in Supplementary Methods. A 323-word stop-word set combined the Harbin Institute of Technology Chinese stop list with a domain-specific stop-word layer comprising templated pedagogical terms (e.g., "教学" [teaching], "讲解" [lecture], "查房" [ward rounds]), generic medical verbs and imaging entities (e.g., "治疗" [treatment], "管理" [management], "CT", "MRI"), and high-frequency clinical narrative scaffolds (e.g., "临床" [clinical], "病历" [medical record], "流程" [workflow], "教学效果" [teaching effect]) added iteratively after early model inspection identified them as topic-uninformative dominators of a residual catch-all topic. The stop list and lexicon were frozen after final BERTopic model selection and before the 200-record LLM-validation sample was drawn; the complete stop list is reproduced in Supplementary Methods. Protected health information (PHI) was removed by a layered 13-rule regular-expression pipeline (PII v3) producing 46,313 placeholder substitutions across the corpus (residency: 34,358 substitutions, dominated by 34,216 healthcare-provider name placeholders; internship: 11,955 substitutions, dominated by 11,730 healthcare-provider name placeholders), with PHI categories (provider name, patient name, medical record number, bed number, staff ID, national ID, mobile, email, long-numeric fallback) replaced by category-specific masked tokens; the full category-level breakdown is in Supplementary Table S8.

BERTopic Topic Modeling

For each system (residency, internship, and the joint residency-plus-internship corpus), we fit a BERTopic pipeline [4]. Document-level vectorization used the BAAI/bge-base-zh-v1.5 768-dimensional Chinese sentence-transformer from the C-Pack family [17] trained with the SBERT framework [16]. We selected BAAI/bge-base-zh-v1.5 as a practical balance between embedding capacity and inference cost on commodity hardware; clinical-pretrained Chinese embeddings (e.g., MedBERT-zh, MC-BERT) and larger general models (e.g., bge-large-zh, multilingual-Qwen) remain candidates for future external sensitivity analyses. The resulting embeddings were reduced with UMAP [14] using n_neighbors = 15, n_components = 5, the cosine metric, min_dist = 0, and random_state = 42, and the reduced vectors were then clustered with HDBSCAN [15] using min_cluster_size = 10 for the residency and joint corpora and 7 for the internship corpus (chosen to reflect its smaller size), with the euclidean metric and cluster_selection_method = "eom". Topic representation used class-based TF-IDF over the jieba-tokenized text [18] within a CountVectorizer that applied the custom stop-word list described above (min_df = 3, max_df = 0.85). Raw BERTopic output was then compressed via BERTopic.reduce_topics(nr_topics = 25) to a parsimonious 24-topic taxonomy per system, with one topic identifier reserved for outliers, and topic centroids in the 768-dimensional embedding space were saved for cross-system mapping. The choice of 24 reduced topics balanced (a) operator interpretability and tabular summarization for a general-medicine audience and (b) empirical coherence, a sensitivity analysis over nr_topics ∈ [15, 20, 24, 30, 40] confirmed that c_v, NPMI, topic diversity, and outlier rate vary within narrow bands across this range (Supplementary Table S15).

Topic Quality, Stability, and Hyperparameter Sensitivity

Three independent quality dimensions were quantified per system. First, topic coherence used the gensim 4.4 CoherenceModel implementation of c_v [9] and NPMI computed against the same jieba-tokenized reference corpus using the top 10 c-TF-IDF words per topic. Second, topic diversity [10] was the proportion of unique tokens within the union of top-10 words across all topics. Third, bootstrap topic stability used five UMAP random seeds (42, 123, 456, 789, 1024). Five seeds were selected as a minimum reproducibility check on the embedding+UMAP+HDBSCAN pipeline (the dominant cost is encoding); within-seed n_topics CV of 3.2-3.4% supports the choice. Pairwise adjusted Rand index (ARI) and normalized mutual information (NMI) summarized stability of cluster assignments across seeds; ARI and NMI were computed including the -1 outlier cluster as a label, treating it as one of the assignment classes. Within-seed n_topics refers to the raw HDBSCAN cluster count before the reduce_topics(nr_topics = 25) reduction step; the post-reduction count of 24 topics per system is fixed by design and is reported separately in Results. As hyperparameter sensitivity analyses, BERTopic was refit on cached embeddings with min_cluster_size ∈ [5, 7, 10, 15, 20, 30] (Supplementary Tables S1-S2) and, separately, with HDBSCAN min_samples ∈ [1, 5, 10, default] crossed with cluster_se-

lection_ epsilon ∈ [0.0, 0.05, 0.1, 0.2] (Supplementary Table S10). As an external calibration check, a gensim Latent Dirichlet Allocation (LDA) baseline (n_topics = 24, passes = 10, iterations = 200, seed = 42) was fit on the same jieba-tokenized stop-listed corpus, and c_v / NPMI / topic-diversity values were compared directly with the BERTopic baseline (Supplementary Table S13).

Multi-LLM Topic Validation

To assess whether the BERTopic taxonomy is interpretable to independent annotators, we adopted a multi-LLM zero-shot encoding protocol consistent with recent methodological recommendations for LLM-based text annotation [7,8,11]. We sampled 200 records (110 residency, 90 internship; stratified across 16 of 24 reduced topics that had at least 10 candidate documents each at character length ≥ 200; the remaining 8 small topics, each with fewer than 10 long-form candidate documents, were under-represented in this validation set, which moderately constrains inter-LLM agreement to the larger 16-topic substructure of the taxonomy; sampling seed = 42). The 200-document sample was drawn from the final reduced 24-topic models after all stop-word-list freezes and topic-reduction decisions; no LLM annotation result was used to modify the BERTopic pipeline or any preprocessing rule. Each record was independently annotated by two commercial LLMs, Codex (GPT-5.2, OpenAI, accessed 2026-05-10) and Gemini 3 (Google DeepMind, accessed 2026-05-10), using identical zero-shot prompts that supplied the 24 reduced topics with their top-10 c-TF-IDF words and one-line glosses (Anthropic Claude Opus 4.7-generated; full prompt in Supplementary Methods §4), and asked the model to return one topic identifier per document or "-1" if no listed topic fit. Inference parameters: temperature = 0 (deterministic), max_tokens = 32 per record, no system-prompt customization; batches were submitted in stratified-sampled order without intra-batch shuffling; outputs were parsed by regex ^-?\d+$ after stripping whitespace and would have been rejected on non-integer output (none observed). Documents were submitted in batches of 20 records each (10 batches per LLM) to avoid context dilution and reduce hallucination risk. Pairwise inter-LLM Cohen's κ on topic identifiers [12] quantified between-model encoding reproducibility, computed both on the full 200-record set and on the subset excluding pairs where either LLM assigned -1 (Supplementary Table S9). We explicitly note that inter-LLM agreement reflects encoding re-

Table 1. Sample characteristics, BERTopic model summary, and bootstrap topic stability across two parallel teaching systems.

Characteristic	Residency	Internship	Joint
Sample size (records, post-cleaning)	2,890	1,921	4,811
Annual records 2022	237	206	443
Annual records 2023	423	319	742
Annual records 2024	397	471	868
Annual records 2025	955	481	1,436
Median character length per record	287	245	273
Departments represented	27	27	27
BERTopic raw n_topics	96	100	159
BERTopic reduced n_topics	24	24	24
Outlier rate (%)	23.4	13.6	16.6
Reduced c_v	0.505	0.548	0.517
Reduced NPMI	-0.030	-0.060	0.002
Reduced topic diversity	0.904	0.867	0.863
5-seed ARI mean ± SD	0.595 ± 0.016	0.674 ± 0.040	–
5-seed NMI mean ± SD	0.898 ± 0.004	0.923 ± 0.009	–
5-seed n_topics mean ± SD	97.0 ± 3.1	96.6 ± 3.3	–

producibility, not agreement with a human gold standard [11], and the topic taxonomy's semantic quality is independently supported by NPMI coherence and bootstrap stability above.

Human Expert Validation

To anchor the multi-LLM encoding-reproducibility statistic to a human gold-standard reference and respond to the reviewer's request for direct human validation, we conducted a 30-record dual-clinician annotation on the top-10 joint-corpus topics. Sample design: three records were drawn per topic for each of the 10 largest joint topics (T0-T9), yielding 30 records in total. For T0-T7 (24 records), the three records per topic were drawn from the original 200-record multi-LLM validation set, allowing direct four-rater comparison with Codex (GPT-5.2) and Gemini 3. For T8 and T9 (6 records), the original 200-record set held fewer than three long-form (character length ≥ 200) candidates, so an additional three records per topic were freshly drawn from the full joint corpus; these short-form records (median character length 13-17) were annotated only by the two human raters because the LLMs were not re-queried on out-of-sample records. Annotators: two board-certified clinical-trainee physicians who are also co-authors of this manuscript (Chujie Chen, MD, surgical resident; Jun Li, MD, attending surgeon at the same institution). Both annotators worked independently and were blinded to the BERTopic-assigned topic identifier, the Codex annotation, and the Gemini annotation; each received a structured packet containing the full Chinese narrative text, the 24-topic option list with English topic labels plus top-10 c-TF-IDF words (identical to the prompt that had been presented to the LLMs; full packet in Supplementary Methods §6), and a one-page Chinese-language decision-rule document. The annotation rule was to assign a single topic identifier 0-23 to each record or "-1" if no listed topic provided a meaningful fit; ties between two equally plausible topics were to be broken in favor of the more specific clinical topic over a generic procedural or didactic topic. Statistical analysis used Cohen's κ on

Table 2. Top 10 largest reduced topics in the joint corpus (n = 4,811).

Topic	Label	Size	Top-6 representative words (English)
0	General clinical reasoning teaching	2,282	reasoning / acute / etiology / clinical manifestations / tumor / drugs
1	Gynecologic-obstetric ultrasound	403	ultrasound / uterus / fetus / gynecology / obstetrics-gynecology / pregnancy
2	Cervical and lumbar spinal disorders	298	cervical spondylosis / fracture / lumbar spine / clinical manifestations / spondylolisthesis / osteoporosis
3	Suturing and wound dressing	167	knot tying / dressing change / wound / suturing / dressing / suture removal
4	Cardiac valvular disease	95	mitral valve / valvular disease / regurgitation / heart / aortic valve / valve
5	Acute appendicitis management	85	acute appendicitis / appendix / appendicitis / surgery / colic / partial
6	Thyroid nodule oncology	80	thyroid / thyroid cancer / postoperative / nodule / ultrasound / nerve
7	Urinary catheterization	76	nasogastric tube / bladder / insertion / urinary catheterization / urinary catheter / sterile
8	Thoracic surgery anatomy	60	thoracic surgery / nodule / surgery / chest / mediastinum / anatomy
9	Emergency-trauma neurosurgery	56	neurosurgery / emergency / surgery / nutrition / craniocerebral / trauma

Note: The Top-6 representative words shown below are English translation glosses of the original Chinese c-TF-IDF tokens (translated by Anthropic Claude Opus 4.7 from BERTopic top-10 c-TF-IDF Chinese tokens). The Chinese original tokens for all 24 topics in each system, alongside their English glosses, are provided in Supplementary Tables S5 (residency) and S6 (internship); readers should consult these for the authoritative Chinese vocabulary that grounded the topic model.

the 30-record full sample for the primary human–human comparison, the 24-record T0-T7 subset for like-for-like comparison with each LLM, and the 6-record T8/T9 fresh-draw subset for human-pair concordance on short out-of-sample records; bootstrap 95% confidence intervals followed the same 1000-iteration, seed = 42 protocol applied above to the LLM-LLM agreement. Two sensitivity versions of each human-LLM κ were computed: including all 24 records (the LLM's "-1" treated as a valid label, mirroring its decision behavior), and excluding records where either rater (human or LLM) returned "-1". Per-topic concordance counted, for each of T0-T7, the proportion of records on which all four raters

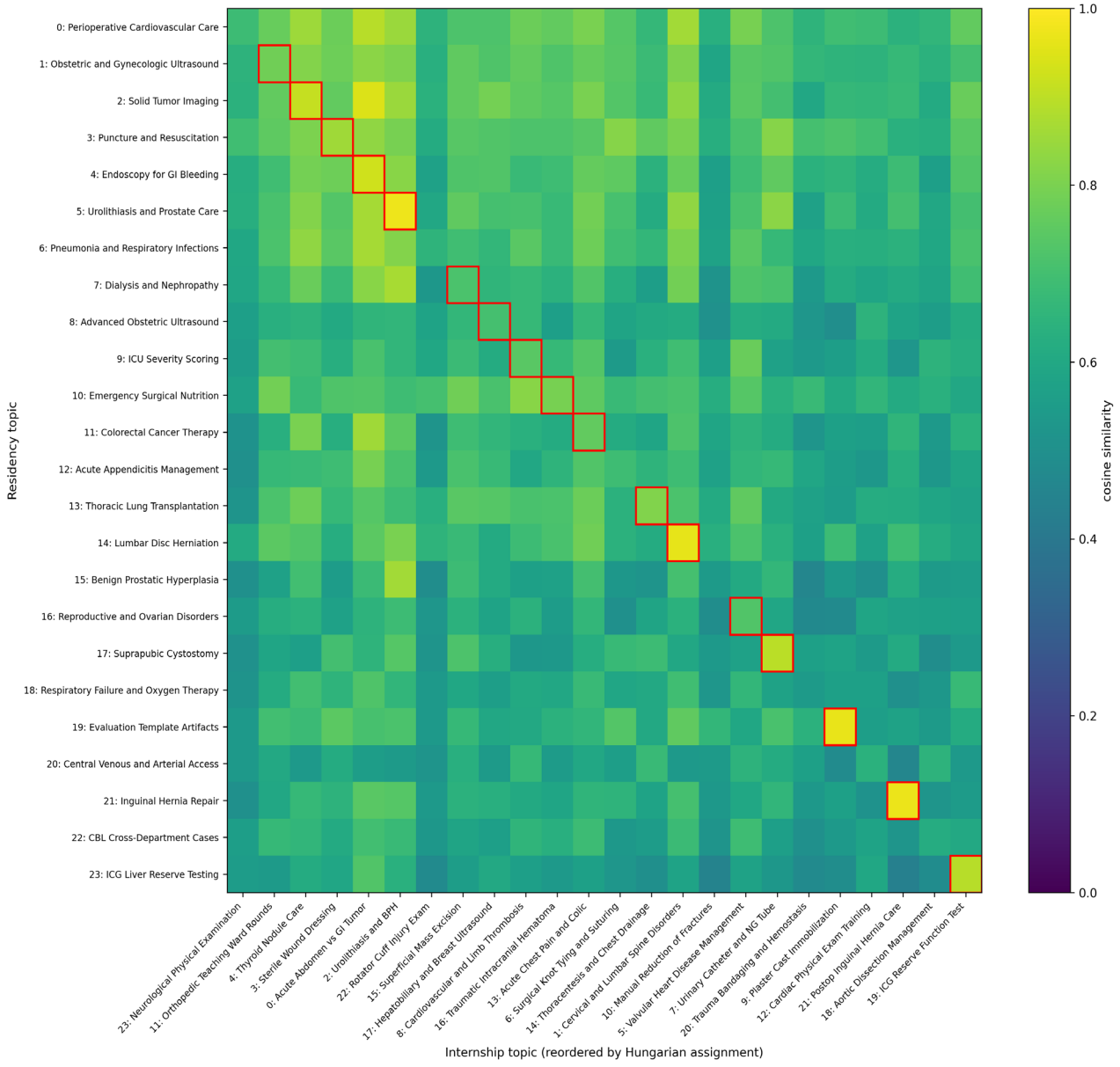

Figure 2. Cosine-similarity heatmap of Hungarian-matched topic pairs between residency 24 reduced topics (rows) and medical internship 24 reduced topics (columns). Internship columns are reordered so that matched pairs lie on the diagonal by Hungarian assignment construction; red boxes mark the 17 pairs reaching cosine ≥ 0.70 (descriptive threshold). The overall 17/24 count did not exceed a 1000-iteration merged-relabel permutation-null expectation (p = 0.76; Supp Table S11), and one apparent high-cosine pair (residency T19 "Evaluation Template Artifacts" ↔ internship T9 "Plaster Cast Immobilization", 0.97) reflects shared template-vocabulary contamination rather than genuine cross-system content. Specific high-cosine pairs at the individual clinical-concept level (e.g., urolithiasis ↔ urolithiasis at 0.97) remain interpretable.

agreed, three of four agreed, or no majority emerged; for T8-T9 the analogous count was the proportion of the three records on which both human raters agreed.

Cross-System Topic Mapping and Temporal Drift

For each pair of (residency topic, internship topic), we computed cosine similarity between topic centroids in the 768-dimensional sentence-embedding space. The Hungarian algorithm [13] (scipy.optimize.linear_sum_assignment) returned the optimal one-to-one matching that maximized the sum of pairwise similarities. To examine temporal drift, we computed each topic's annual share of within-year teaching activities (2022, 2023, 2024, 2025) per system and identified the top 5 topics with the largest share gain and the largest share loss over the four-year window.

Software and Statistical Analysis

All analyses were performed in Python 3.13 using BERTopic 0.17, sentence-transformers 5.1, jieba 0.42, scipy 1.14, gensim 4.4 (CoherenceModel implementation following Newman and colleagues [27] and Aletras and Stevenson [26]), scikit-learn 1.6 [19], and matplotlib 3.10. Pairwise inter-LLM agreement followed Cohen's original formulation [21]; bootstrap-stability ARI followed Hubert and Arabie [22], and NMI followed Strehl and Ghosh [23]. The gensim implementation framework is described in Rehurek and Sojka [20]. Conceptual framing of LLM-based annotation as a complement (not replacement) for traditional supervised approaches follows recent commentary on machine-learning–statistical-learning integration in clinical research [24]. N-gram-based coherence concepts trace to early work in distributional semantics [25]. The full analysis code, reproducibility manifest, and de-identified text corpus are available from the corresponding author on reasonable requests, subject to institutional data governance restrictions on the underlying CCMTV teaching activity records. The IRB-approved analytic plan, model artifacts, derived topic-quality reports, prompt templates, the complete 323-word stop list, the 50-term lexicon, and the 200-document PHI-substituted validation set are deposited as Supplementary Material.

Results

Sample

The final corpus comprised 4,811 teaching activity records spanning 2022-2025, with annual counts of 443 (2022), 742 (2023), 868 (2024), and 1,436 (2025) (Table 1, top panel). Residency contributed 2,890 records (60%) and medical internship 1,921 (40%). Median per-record character length was 287 (residency) and 245 (internship). Records spanned 27 clinical departments, with the largest contributions from internal medicine, surgery, gynecology–obstetrics, and emergency medicine subspecialties.

Topic Landscape

After topic reduction to 24 topics per system,

Table 3. Top 10 highest-similarity cross-system topic pairs.

#	Residency topic	Internship topic	Cosine similarity
1	T5: Urolithiasis and Prostate Care	T2: Urolithiasis and BPH	0.974
2	T21: Inguinal Hernia Repair	T21: Postop Inguinal Hernia Care	0.970
3	T19: Evaluation Template Artifacts	T9: Plaster Cast Immobilization	0.968
4	T14: Lumbar Disc Herniation	T1: Cervical and Lumbar Spine Disorders	0.962
5	T4: Endoscopy for GI Bleeding	T0: Acute Abdomen vs GI Tumor	0.930
6	T2: Solid Tumor Imaging	T4: Thyroid Nodule Care	0.914
7	T17: Suprapubic Cystostomy	T7: Urinary Catheter and NG Tube	0.895
8	T23: ICG Liver Reserve Testing	T19: ICG Reserve Function Test	0.892
9	T3: Puncture and Resuscitation	T3: Sterile Wound Dressing	0.858
10	T13: Thoracic Lung Transplantation	T14: Thoracentesis and Chest Drainage	0.811

Note: Likely spurious, shared template-vocabulary contamination, not curriculum content. The full 24-pair table with permutation-null p-values is in Supplementary Table S11.

the residency model reached c_v = 0.505 (mean of 24 reduced topics; range 0.30-0.62), NPMI = -0.030, and topic diversity = 0.904; internship reached c_v = 0.548, NPMI = -0.060, diversity = 0.867; and the joint corpus reached c_v = 0.517, NPMI = 0.002, diversity = 0.863. Outlier rates were 23.4% (residency), 13.6% (internship), and 16.6% (joint). Recommended interpretive ranges treat c_v ≥ 0.55 as good and 0.40 - 0.55 as acceptable [9], and a topic-diversity threshold of > 0.70 is regarded as practically acceptable [10]; all reduced models met these thresholds. Topic labels generated by an independent third LLM (Anthropic Claude Opus 4.7) for each of the 72 reduced topics demonstrated strong alignment between top-10 words and assigned labels, with 71 of 72 topics showing direct semantic correspondence and one topic ("residual evaluation template tokens") flagged as residual.

The 24 joint-corpus topics spanned the major operational categories of clinical teaching: specialty-specific topics (e.g., gynecologic ultrasound; cervical/lumbar disorder; cardiac valvular disease), procedural skills topics (e.g., suturing/wound dressing; urinary catheterization; thoracic-tube drainage), and structural/integrative teaching topics (e.g., clinical reasoning, multidisciplinary case-based learning). Table 2 lists the 10 largest joint-corpus topics with size and labels; the full set of 24 reduced-topic labels with top-10 c-TF-IDF words for each system is provided in Supplementary Tables S5–S6.

Stability and Hyperparameter Sensitivity

Five-seed UMAP bootstrap revealed substantial cluster-assignment stability: residency ARI = 0.595 ± 0.016, NMI = 0.898 ± 0.004, n_topics = 97.0 ± 3.1 raw topics (CV 3.2%); internship ARI = 0.674 ± 0.040, NMI = 0.923 ± 0.009, n_topics = 96.6 ± 3.3 (CV 3.4%) (Table 1, middle panel). Across the min_cluster_size grid (K = 5, 7, 10, 15, 20, 30), residency c_v varied within a narrow band (0.513-0.586) and internship c_v varied within 0.538-0.552 (Figure 4; Supplementary Tables S1–S2). HDBSCAN sensitivity to min_ samples and cluster_selection_ epsilon (Supplementary Table S10) showed raw topic counts spanning 89-131 (residency) and 66-119 (internship), with outlier rates 6.0-17.3% (residency) and 2.3-18.1% (internship); the chosen baseline (residency K = 10, internship K = 7) maximized NPMI for residency and lay in the middle of the K range for internship, supporting the choice as a defensible default rather than a c_v-maximizing point estimate. An external LDA baseline (gensim, n_topics = 24, same corpus and tokenization) reached c_v 0.439/0.508 and topic diversity 0.546 / 0.504 (residency / internship), confirming that BERTopic c_v exceeds LDA baseline c_v on the same corpus and that BERTopic produces substantially higher topic diversity (Supplementary Table S13). LDA NPMI (+0.011/+0.044) was higher than BERTopic NPMI (-0.030/-0.060), a known property of NPMI on c-TF-IDF-tokenized embedding-based topics; we report both metrics.

Multi-LLM Topic Validation

Two independent commercial LLMs (Codex GPT-5.2 and Gemini 3) returned full annotations for all 200 stratified-sampled records (200/200 entries per LLM). Inter-LLM raw agreement was 72.5% (145 of 200 documents assigned identical topic identifiers), and Cohen's κ on topic identifiers was 0.709 (1000-iteration bootstrap 95% CI 0.645-0.776), corresponding to the substantial range of the Landis and Koch convention [12] (κ ≥ 0.61), though we note that Landis-Koch labels are conventional bookkeeping rather than thresholds for validation. Codex assigned the "no fit" residual identifier (-1) to 21.5% of records (43 of 200), reflecting a more conservative annotation behavior; Gemini used "-1" for only 3.0% of records (6 of 200), preferring to assign most records to their closest available topic. Excluding pairs where either LLM assigned -1 (n = 157), Cohen's κ rose to 0.878 (95% CI 0.823-0.925) and raw agreement to 88.5% (Supplementary Table S9); a 24 × 24 per-topic confusion matrix is provided in Supplementary Table S9b, showing diagonal dominance with no systematic mis-assignment pattern. These results indicate that residual-class disagreements are the dominant driver of inter-LLM dispersion within the substantive 24-topic taxonomy. Among the 145 records on which the two LLMs concurred, the 24-topic taxonomy was approximately uniformly used (largest topic concordance counts: T0 "general clinical reasoning teaching" 18 records; T6 "thyroid nodule oncology" 8 records; T2 "cervical/lumbar spinal disorders" 7 records; complete pairwise concordance counts in Supplementary Table S7). The substantial inter-LLM encoding-reproducibility supports the BERTopic taxonomy as semantically encoding-reproducible across two independent commercial language models, while we explicitly retain the conceptual distinction between LLM-vs-LLM agreement and human expert validation; the latter is reported in the next subsection.

Human Expert Validation

For the 30-record dual-clinician annotation across the top-10 joint topics, Cohen's κ between the two board-certified human annotators was 0.782 (1000-iteration bootstrap 95% CI 0.603-0.925; raw agreement 80.0%) on the full 30 records, and 0.770 (95% CI 0.580-0.906; raw agreement 79.2%) on the 24-record T0-T7 subset matched to the multi-LLM validation set. On the 6-record T8/T9 fresh-draw subset, human-pair κ was 0.786 (95% CI 0.357-1.000; raw agreement 83.3%), confirming that human concordance on short out-of-sample records was comparable to the LLM-eligible subset, although the small N produces a wide CI. Both human-pair values fall within the substantial-to-near-perfect band of the Landis and Koch convention [12] and overlap the inter-LLM κ reported above (κ = 0.709 on the full 200-record set), placing the two-human gold-standard references squarely within the same agreement band as the two-LLM annotators. Human-versus-LLM κ on the 24-record T0-T7 subset, with the LLM's "-1" treated as a valid label, ranged from 0.632 (Chujie-Codex, 95% CI 0.420-0.812) to 0.816 (Jun Li-Gemini, 95% CI 0.632-0.952). Excluding records where either rater returned "-1", human-LLM κ rose to 0.807-0.936 (raw agreement 0.826-0.944), and the LLM-LLM subset agreement reached κ = 1.000 (raw agreement 1.000) on the 18 of 24 records on which both LLMs assigned a substantive (non–"-1") label, paralleling the pattern reported in the multi-LLM analysis above where excluding the LLM's residual category sharply raised inter-LLM agreement. Per-topic four-rater concordance (Supplementary Table S17) showed that on the four topics T3 (suturing/wound dressing), T5 (urolithiasis), T6 (thyroid nodule oncology), and T7 (urinary catheterization), all four raters agreed on all three records (12 of 12 records all-agree); on topics T0 (general clinical reasoning) and T4 (cardiac valvular disease), most records achieved majority (≥ 3 of 4) agreement; and on topics T1 (gynecologic ultrasound) and T2 (cervical/lumbar spinal disorders), four-rater agreement was lower, with both human annotators converging on T14 (cervical/lumbar spinal disorders) rather than the BERTopic-assigned T2 on one record, flagging a possible boundary mis-assignment between two adjacent musculoskeletal topics in the joint taxonomy. On the 6 T8/T9 fresh-draw records, the two human raters agreed on 5 of 6 records, supporting the interpretability of these two smaller joint topics even for short narratives. Taken together, the human expert validation anchors the multi-LLM encoding-reproducibility statistic to a human gold-standard reference at the same agreement band, identifies one specific topic boundary (T2/T14) as a candidate for taxonomy refinement in future re-fits, and supports the BERTopic-plus-multi-LLM workflow as semantically interpretable to board-certified clinical educators without specialist NLP training.

Cross-System Topic Mapping

Hungarian-matched topic pairs (residency 24 ↔ internship 24) achieved cosine similarity (Figure 2; Table 3). Of 24 paired topics, 17 (71%) reached cosine ≥ 0.70 (high-similarity, descriptive), 7 (29%) reached 0.50-0.70 (moderate), and none fell below 0.50. We caution that under a 1000-iteration merged-relabel permutation null model (pooling residency and internship centroids and randomly re-partitioning), the observed mean cosine (0.783) was not significantly above the null mean (0.776; $p = 0.32$), and the observed N ≥ 0.70 (17) was not above the null mean (17.5; $p = 0.76$; Supplementary Table S11). The population-level "71% overlap" statistic should therefore be read as descriptive rather than as evidence of dual-system-specific shared curriculum structure; specific high-cosine pairs (e.g., urolithiasis ↔ urolithiasis at 0.97) remain interpretable at the individual clinical-concept level. The strongest cross-system correspondences (top-5 cosine ≥ 0.93) all reflected highly specialized clinical content: urolithiasis and benign prostatic enlargement (residency T5 ↔ internship T2, cosine 0.974), inguinal hernia repair (T21 ↔ T21, 0.970), cervical and lumbar spinal disorders (T14 ↔ T1, 0.962), gastrointestinal bleeding and endoscopy (T4 ↔ T0, 0.930), and solid-tumor differential imaging (T2 ↔ T4, 0.914). Weakest matches (cosine 0.55–0.65) corresponded to system-specific content: respiratory failure and oxygen therapy (residency-specific) was forced-matched to wound-bandaging-and-hemostasis (internship-specific) because the Hungarian assignment requires a one-to-one matching even when no genuine cross-system counterpart exists.

Temporal Drift, 2022–2025

Topic share by year (Figure 3; Supplementary Tables S3-S4; Supplementary Figures S1-S2; non-outlier per-year sample sizes: residency 168/312/322/770; internship 181/261/397/396 for 2022/2023/2024/2025, with 95% Wilson confidence intervals shaded in figures) revealed substantial topic-share change in both systems. Endpoint deltas (2025 share minus 2022

share, not slope) drove the top-5 rising/falling ranking. To formally test year effects, we fit a per-topic-per-system binomial GLM with year (2022-2025) as integer predictor (Supplementary Table S14) and applied Benjamini-Hochberg FDR over the 24 topics within each system: 17 of 24 residency topics (71%) and 15 of 24 internship topics (63%) showed significant year trends at FDR < 0.05. The largest year effects (odds ratios per year far from 1.0) corresponded to the topics highlighted in the endpoint-delta ranking, confirming that the descriptive pattern is not driven solely by 2022-vs-2025 noise. In the residency corpus, the largest endpoint gains were "perioperative cardiovascular management" (+28.7 percentage points; from 4.2% in 2022 to 32.9% in 2025), "clinical puncture and emergency procedures" (+13.2 pp; 0% to 13.2%), and "solid-tumor differential imaging" (+12.2 pp; 6.3% to 18.5%); the largest losses were "obstetric ultrasound" (-28.0 pp; 35.0% to 7.0%), "advanced obstetric ultrasound" (-11.0 pp), and "emergency surgical nutrition support" (-10.3 pp). In the internship corpus, the largest gains were "acute abdomen and oncology differential" (+23.3 pp; 5.8% to 29.1%), "wound dressing and asepsis" (+8.6 pp), and "thyroid nodule oncology" (+4.7 pp); the largest losses were "cervical/lumbar disorders" (-7.1 pp), "orthopedic teaching rounds" (-6.7 pp), and "thoracic-tube drainage" (-5.1 pp). Annual outlier rates were stable within each system (residency 19-29%; internship 12-18%; Supplementary Table S12) without monotonic trend, suggesting BERTopic outlier reassignment is not driving the observed topic-share drift. The convergent observation across both systems was the rise of broad clinical-reasoning topics ("perioperative cardiovascular" and "acute abdomen / oncology differential") at the expense of specialty-specific procedural topics, a topic-share shift compatible with a possible curricular reorientation from procedure-focused demonstrations toward integrative clinical-reasoning teaching but interpretable only as descriptive change in the recorded corpus pending external validation.

Discussion

Principal Findings

In a four-year single-center corpus of 4,811 free-text teaching narratives from a tertiary Chinese teaching hospital, BERTopic recovered semantically coherent 24-topic taxonomies for both residency and medical internship training, with model coherence c_v 0.51-0.55, topic diversity 0.86-0.90, outlier rates of 14–23%, and substantial cross-seed stability (ARI 0.60-0.67; NMI ≥ 0.90). A multi-LLM zero-shot validation protocol using two independent commercial LLMs (Codex GPT-5.2 and Gemini 3) produced substantial inter-LLM agreement on topic assignment (Cohen's κ = 0.709, raw agreement 72.5%) for 200 stratified-sampled records. Cross-system topic mapping in 768-dimensional embedding space showed that 71% of topic pairs reached cosine similarity ≥ 0.70, indicating that the two parallel training systems share the bulk of their clinical-teaching topic structure while preserving a minority of system-specific content. Over four years, both systems showed convergent drift toward broader clinical reasoning topics at the expense of specialty-specific procedural topics.

Methodological Contribution: Inter-LLM Encoding Reproducibility as a Scalable Complement

The principal methodological contribution is the demonstration that two off-the-shelf commercial LLMs can independently annotate the same 200-document encoding-reproducibility set under identical zero-shot prompts and reach substantial inter-LLM agreement on a 24-class topic-assignment task. The Cohen's κ of 0.709 (full set including -1 residuals; 0.878 excluding -1 on either side, Supplementary Table S9) falls within the substantial-to-near-perfect range of the Landis and Koch reference scale [12], is achievable in approximately 7 minutes of LLM API time per model (40-45 seconds per 20-document batch), and offers a scalable complement to the multi-week effort that would be required to recruit and train multiple human expert annotators. Crucially, both LLMs reached this agreement without specialist medical-education prompting and using only the BERTopic-generated top-10 words and one-line glosses (Anthropic Claude Opus 4.7-generated; see Supplementary Methods) as topic definitions, suggesting that the topic taxonomy is interpretable without specialist priors. We deliberately frame this result as inter-LLM encoding reproducibility, not as agreement with a human gold standard, consistent with prior cautions about the risk of correlated biases across LLMs trained on overlapping pre-training corpora [11]; we therefore report inter-LLM agreement only alongside corpus-level coherence (NPMI) and seed-bootstrap stability, which are independent of LLM annotation behavior. We also report the divergent residual-assignment behavior of the two LLMs (Codex 21.5% "no fit" vs Gemini 3.0%) as a transparent observa-

tion: a strict-match criterion lowers agreement only modestly, and the substantive topic-assignment correlation remains robust (Supplementary Table S9). To anchor this multi-LLM encoding-reproducibility statistic to a human gold-standard reference, we additionally conducted a 30-record dual-clinician validation on the top-10 joint topics (Results §"Human Expert Validation"). The two board-certified human annotators reached κ = 0.782 (95% CI 0.603-0.925) on the full sample and κ = 0.770 on the LLM-eligible 24-record subset, and human-LLM κ reached 0.807-0.936 on the same subset after excluding the LLM's "-1" no-fit residual, values that fall in the same substantial-to-near-perfect agreement band as the LLM-LLM statistic, supporting the interpretation that the multi-LLM encoding-reproducibility figure is not an artifact of correlated LLM biases but reflects topic-assignment agreement that human expert raters also independently produce. The human validation additionally identified one specific cross-topic boundary

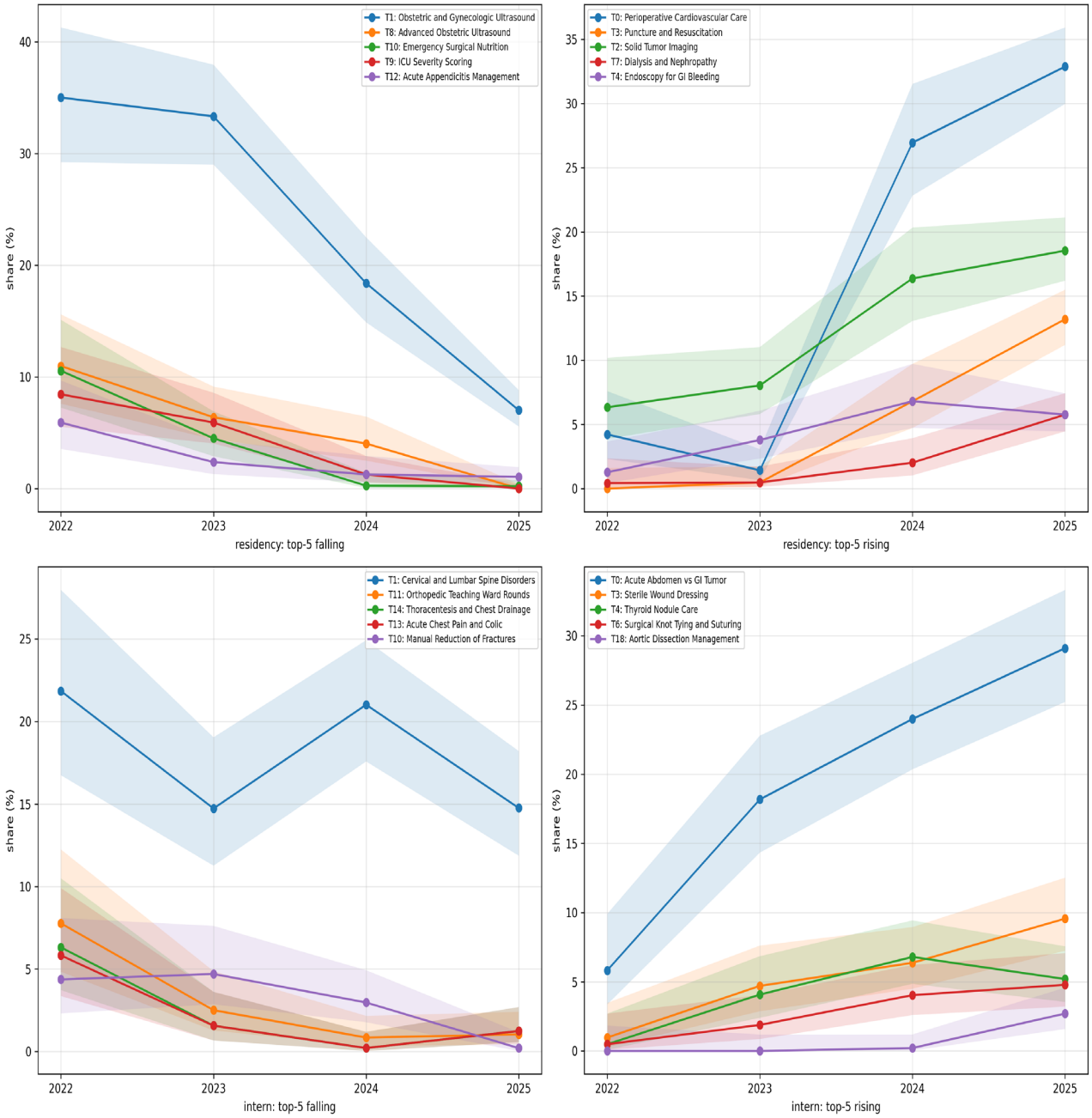

Figure 3. Annual topic share of the 5 fastest-rising and 5 fastest-falling topics in each system between 2022 and 2025, illustrating the convergent shift toward general clinical reasoning topics.

(between T2 cervical/lumbar spinal disorders and T14 cervical/lumbar spinal disorders) as a candidate for taxonomy refinement in future re-fits, a substantive correction that the LLM-only protocol did not surface and that illustrates the complementary, not redundant, value of a small human validation layer attached to a larger LLM encoding-reproducibility check.

Substantive Insight: 71% Cross-System Topic Overlap and Curricular Convergence

The cross-system mapping result, 17 of 24 topic pairs with cosine similarity ≥ 0.70 (mean 0.783), is consistent with the residency and medical internship programs at this institution teaching a substantially shared set of clinical topics. We caution, however, that a 1000-iteration merged-relabel permutation null model (Supplementary Table S11) produced null mean cosine of 0.776 (p = 0.32 vs observed) and null mean N ≥ 0.70 of 17.5 (p = 0.76 vs observed 17), indicating that part of the high cross-system overlap reflects the homogeneous density of the BGE embedding space across all medical-teaching narratives rather than a dual-system-specific signal. Conversely, individual high-cosine pairs at the specific clinical-concept level remain interpretable: the strongest cross-system pairs (cosine ≥ 0.93) all map specialized clinical content (urolithiasis, hernia repair, spinal disorders, GI bleeding, oncologic imaging) and one apparent high-cosine pair (residency T19 "Evaluation Template Artifacts" ↔ internship T9 "Plaster Cast Immobilization", cosine 0.97) reflects shared template-vocabulary contamination rather than genuine cross-system shared curriculum content (Supplementary Table S11), suggesting that core clinical teaching content is taught congruently to learners at both training stages while a residual artifact pair is flagged as spurious. The 7 of 24 weak matches (cosine < 0.70) correspond to system-specific content: residency-specific topics include respiratory failure/oxygen therapy and case-based-learning cross-departmental sessions; internship-specific topics include orthopedic procedural skills and shoulder pathology. This pattern is consistent with the conventional positioning of the medical internship as a broader "general clinical exposure" stage and residency as a more specialized "deepening" stage, but to our knowledge no prior study has quantified this relationship at the level of topic structure.

Convergent Drift Toward Clinical Reasoning Topics

The temporal topic-share analysis revealed a convergent pattern: the broadest "clinical reasoning" topic in each system (residency: perioperative cardiovascular management; internship: acute abdomen and oncology differential) showed the largest gain in topic share between 2022 and 2025 (+28.7 pp residency, +23.3 pp internship), while specialty-specific procedural topics (obstetric ultrasound in residency; orthopedic procedures in internship) showed the largest decline. Several caveats require explicit discussion. First, broad-content topics may attract assignment of newly diverse 2025 teaching content as a consequence of c-TF-IDF representation rather than genuine curricular intent, a known property of BERTopic that should be interpreted as a hypothesis rather than as proven curricular intent. Second, the four-year window is short and does not separate cyclic from secular trends. Third, the documented activity counts grew substantially over the four years (2022 n = 443 to 2025 n = 1,436); the large 2025 sample could provide more nuanced topic discrimination not available in earlier years, and platform-adoption / documentation-practice changes may contribute to apparent topic-share shifts (see Limitations). Annual outlier rates were stable within each system (residen-

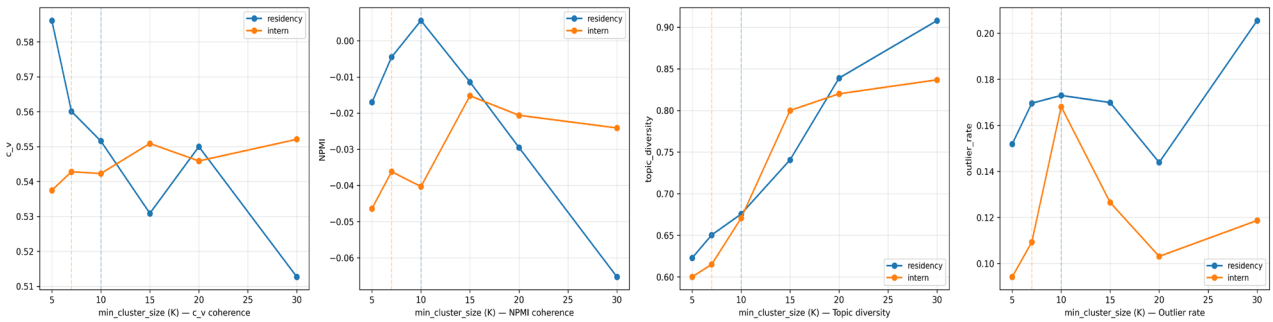

Figure 4. BERTopic min_cluster_size grid sensitivity: c_v, NPMI, topic diversity, and outlier rate as functions of min_cluster_size [5, 7, 10, 15, 20, 30] for residency and medical internship.

cy 19-29%; internship 12-18%; Supplementary Table S12) without monotonic trend, suggesting BERTopic outlier reassignment is not driving the observed topic-share drift. We therefore present the temporal topic-share findings as descriptive and hypothesis-generating, suggesting a possible topic-share shift toward integrative clinical-reasoning teaching that warrants prospective monitoring rather than as a tested causal claim.

Comparison with Prior Work

To our knowledge, few studies have combined BERTopic with a multi-LLM semantic-encoding-reproducibility check for Chinese-language clinical-teaching narratives. Prior natural-language-processing applications in medical education (e.g., NLP-based evaluation of an international medical e-learning course [3]) have typically used English-language reflective writing or program-evaluation comments, an approach that limits both throughput and replication across institutions when applied to Chinese-language operational teaching records. Recent BERTopic applications in biomedical text [5,6] have focused on PubMed abstracts or discharge summaries; their ground-truth labels (MeSH terms; ICD codes) provide structured validation that does not exist for free-text teaching narratives. Our inter-LLM encoding-reproducibility framework offers a portable complement to manual human validation when no structured label set exists.

Implications for Curriculum Monitoring

For institutions operating CCMTV-style electronic teaching platforms, the workflow demonstrated here (BERTopic with multi-LLM zero-shot encoding-reproducibility checking, complemented by a small human expert validation as reported in Methods §"Human Expert Validation") can be re-fit annually or per academic semester to monitor topic distribution and describe potential topic-share changes. The 24-topic taxonomy generated here is institution-specific and not generalizable as a fixed ontology, but the methodology may be adapted by other institutions operating similar electronic teaching management platforms, contingent on local clinical specialty mix, language, and platform-adoption maturity. We identify four concrete use cases for program directors, education leaders, and curriculum committees, and we describe how the output can be integrated into existing program-review workflows.

First, annual or per-semester topic-distribution monitoring: re-fitting the BERTopic pipeline on the rolling teaching-narrative corpus on an annual or per-semester cadence produces standard outputs (topic counts, topic-share trends, top-10 representative words per topic) that can be embedded directly as quantitative exhibits in an institution's annual residency-program self-assessment report or in periodic accreditation submissions, alongside more traditional metrics such as case-log volumes and trainee evaluation scores. Year-over-year topic-share deltas (Figure 3 in the present manuscript) operationalize the qualitative claim of "what we taught this year" into a reproducible quantitative substrate. Second, detection of topic redundancy and curriculum gaps: small, semantically adjacent topics in the catch-all distribution can flag content duplication across rotations or specialty subgroups (for example, several near-duplicate "suturing technique" topics across departments may indicate redundant teaching that could be consolidated into a shared skills-lab session), while the residual outlier (-1) bucket can flag teaching narratives that do not map cleanly to any current topic and may signal under-covered emerging content (for example, post-pandemic shifts in infection control teaching) for curriculum-committee review. Third, cross-program coordination signals: the Hungarian-matched cross-system comparison (Figure 2) gives a curriculum committee a quantitative basis to identify topics taught congruently across residency and internship, supporting coordinated content sequencing, versus topics taught only in one system, which may either reflect intended specialization or unintended divergence requiring deliberate adjustment. The residency-specific "respiratory failure and oxygen therapy" and the internship-specific "wound bandaging and hemostasis" in our corpus exemplify expected stage-appropriate divergence, whereas an unexpected mismatch in a foundational topic (e.g., basic life support) would warrant program-director attention. Fourth, integration with existing program-review workflows: the pipeline produces tabular outputs that drop into existing program-review templates without requiring any new committee process; it complements rather than replaces expert clinical-educator judgement and the standard Kirkpatrick evaluation hierarchy (i.e., this is Level 0 documentation monitoring, not Level 1 reaction nor Level 2/3/4 learning, behavior, or outcome). Two practical considerations bound these use cases. First, the labor cost of one

re-fit on commodity hardware is approximately one engineer-day per system once the pipeline is set up, which is compatible with annual but not weekly cadence. Second, year-over-year topic-share changes must be interpreted alongside known platform-adoption and documentation-practice trends (see Limitations §7); a curriculum committee should not treat a single year's topic-share change as evidence of curricular intent without external context from program leadership.

Limitations

Several limitations warrant acknowledgment. First, our topic encoding-reproducibility checking relied on multi-LLM agreement and corpus-level coherence rather than expert human gold-standard coding. Although recent methodological work shows that multi-LLM voting can approximate expert reliability for routine annotation [7,8], inter-LLM agreement does not directly equal LLM-vs-human agreement and may inflate apparent reliability if the two models share systematic biases derived from overlapping pre-training corpora [11]. We mitigated this by reporting NPMI coherence, bootstrap stability, and an LDA baseline (Supp Table S13) as independent quality indicators, and by executing a 30-record dual-clinician validation across the top-10 joint topics (Methods §"Human Expert Validation"; Results §"Human Expert Validation") in which two board-certified human annotators reached $\kappa = 0.782$ (95% CI 0.603-0.925) and human-LLM agreement reached $\kappa = 0.807$-0.936 on the matched 24-record subset after excluding the LLM's "-1" no-fit residual, placing human-LLM agreement in the same substantial-to-near-perfect band as the multi-LLM encoding-reproducibility statistic. A no-gloss-prompt sensitivity (giving the LLMs only Chinese top-10 c-TF-IDF words, withholding Claude Opus 4.7-generated English glosses) was not run in this baseline analysis and is identified as a deferred robustness check; together with a double bilingual expert review of the 24 topic labels by two independent medical educators, these are the most informative next-step robustness analyses. Second, the data come from a single tertiary teaching hospital in southern China, and the 24-topic taxonomy reflects the local clinical specialty mix and patient case-mix; the specific topic labels (e.g., the prominence of urolithiasis, gynecologic ultrasound, and orthopaedic content) are therefore not directly transferable as a fixed ontology to other institutions. We caution that the methodology (BERTopic + UMAP + HDBSCAN + multi-LLM with human expert zero-shot validation + Hungarian cross-system mapping with permutation-null sanity check + per-topic binomial GLM with BH-FDR) is, in contrast, platform-agnostic and re-deployable on any institution's free-text teaching-narrative corpus subject to a re-fit of the BERTopic model on the local corpus. Institutions operating on other electronic teaching platforms (e.g., non-CCMTV vendor systems, in-house EMR-integrated teaching modules) can run the same pipeline once teaching narratives are exported to plain text, but several local adaptations are required: (a) the PHI substitution layer must be re-calibrated to local naming conventions, identifier formats, and any institution-specific structured fields; (b) the choice of Chinese sentence-transformer (BAAI/bge-base-zh-v1.5) should be reconsidered for non-Chinese-language corpora, a multilingual or English-pretrained embedding (e.g., mpnet-base, multilingual-e5) may be more appropriate; (c) the domain-specific stop list and jieba lexicon must be re-curated for the local clinical vocabulary; and (d) the topic-reduction parameter (nr_topics = 25) should be re-tuned to the local corpus size and specialty breadth. On commodity hardware (single workstation with mid-range GPU), the full pipeline re-fit takes approximately one engineer-day per system once the operator is familiar with Python and the BERTopic API. Pre-registered multi-centre replications using a shared protocol and shared evaluation rubric would be the most informative next-step external validation, and we have prospectively published the analysis scripts and prompt templates as Supplementary Materials to facilitate such replication. Third, BERTopic on corpora of this size is sensitive to UMAP and HDBSCAN hyperparameter choices; we mitigated this through grid searches over min_cluster_size, min_samples × cluster_selection_epsilon (Supp Table S10), and 5-seed bootstrap stability assessment. Fourth, the temporal topic-share findings are descriptive over four years of growing activity counts and require multi-year follow-up to distinguish curricular intent from sample-size effects. Fifth, our analysis is restricted to the topic-distribution layer of teaching narratives (a descriptive monitoring of teaching-content recording, not Kirkpatrick Level 1 reaction or higher levels) and does not measure learner knowledge, behavior, or patient outcomes; the framework is complementary to, not a substitute for, outcome-oriented program evaluation. Sixth, English topic labels were generated by an in-

dependent third LLM (Anthropic Claude Opus 4.7) from BERTopic top-10 c-TF-IDF words and reviewed by one of the authors (CC); we cannot exclude residual translation drift where ambiguous Chinese topics were rendered as cleaner-sounding English names. Chinese top-10 words and full Chinese topic labels are provided in Supplementary Tables S5–S6 to permit independent verification by a Chinese-speaking reader, and a no-gloss sensitivity prompt was not run as part of this baseline analysis. Seventh, documented teaching-activity counts grew substantially over the four years (443 in 2022 to 1,436 in 2025), which may reflect changes in CCMTV adoption and documentation practice rather than true curricular evolution; observed topic-share trajectories should be interpreted as topic-distribution shifts in the recorded corpus rather than as direct evidence of curricular intent. Eighth, the cross-system topic-overlap finding (mean cosine 0.783; 17 of 24 pairs ≥ 0.70) was not significantly higher than a 1000-iteration merged-relabel permutation null (Supplementary Table S11), indicating that the high overlap partly reflects the homogeneous density of the BGE embedding space across all medical-teaching narratives; specific high-cosine pairs (e.g., urolithiasis ↔ urolithiasis, 0.97) remain interpretable at the individual clinical-concept level but the population-level "71% overlap" claim should be read as descriptive rather than as evidence of dual-system-specific structure. Ninth, a formal residual-PHI audit conducted on a 100-record stratified sample (50 residency + 50 internship; Supp Table S16) found no unmasked national-ID, mobile, or email patterns; conservative regex screening of name-like and MRN-like character sequences produced upper-bound flag rates that on inspection consisted predominantly of clinical noun phrases (e.g., disease names, dose figures) rather than true residual identifiers, although a second-researcher visual review is recommended for any subsequent text release. Future multi-center work, longitudinal monitoring of curriculum-policy alignment, and human spot-check anchoring would strengthen these inferences.

Conclusion

BERTopic with multi-LLM zero-shot inter-LLM encoding-reproducibility checking provides a scalable, descriptive framework for monitoring the topic structure of clinical teaching activities at scale. Applied to a four-year, dual-system single-center corpus of 4,811 teaching narratives, the framework recovered coherent and stable 24-topic taxonomies, demonstrated substantial inter-LLM annotation agreement (κ = 0.709 full; 0.878 excluding the -1 no-fit residual) anchored by a 30-record dual-clinician validation in which the two board-certified human annotators reached κ = 0.782 (95% CI 0.603-0.925) and human-LLM agreement reached κ = 0.807-0.936 after excluding the LLM's "-1" residual, characterized cross-system topic correspondences (mean cosine 0.783; 17 of 24 pairs ≥ 0.70, with permutation-null caveats noted), and described a convergent four-year topic-share change toward integrative clinical-reasoning topics in both systems. The framework may be adapted by other institutions operating similar electronic teaching management platforms, contingent on local clinical specialty mix, language, and platform-adoption maturity.

Declarations

Ethics Approval and Consent to Participate

The study was conducted in accordance with the Declaration of Helsinki. The Ethics Committee of the Seventh Affiliated Hospital, Sun Yat-sen University, approved the protocol covering both residency and medical internship teaching activity data extracted from the CCMTV electronic platform (approval number: KY-2026-111-01, approved 2026). Because this study used routinely collected administrative teaching activity records that were de-identified prior to analysis (46,313 protected health information placeholder substitutions across the corpus; full category-level breakdown in Supplementary Table S8), the requirement for informed consent from individual faculty and trainees was waived by the same Ethics Committee.

Consent for Publication: Not applicable. No individual identifying information appears in this manuscript or its supplementary material.

Availability of Data and Materials*:* The aggregated topic-quality reports, topic labels (Chinese and English), cross-system mapping summaries, temporal-drift summaries, prompt templates, complete stop-word list, custom lexicon, and the de-identified 200-document validation set are available on reasonable request from the corresponding author, subject to institutional data governance restrictions on the underlying primary teaching activity records.

Funding Support: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest: The authors declare no competing interests.

Authors' Contributions: Chujie Chen: conceptualization, methodology, software, formal analysis, data curation, writing, original draft, visualization. Jun Li: conceptualization, supervision, writing, review & editing, project administration. All authors approved the final manuscript.

Acknowledgements: We thank the clinical teaching faculty at the Seventh Affiliated Hospital, Sun Yat-sen University, whose recorded teaching activities form the corpus analyzed in this study.

The authors used DeepSeek (DeepSeek AI; https://deepseek.com) to assist with text translation between Chinese and English during manuscript preparation; the authors are solely responsible for the scientific content and accuracy of the manuscript. All analytic uses of large language models, Codex (GPT-5.2) and Gemini 3 as independent zero-shot annotators for the multi-LLM topic-validation set, and Anthropic Claude Opus 4.7 for one-line topic-label gloss generation, are described in Methods §"Multi-LLM Topic Validation" and Supplementary Methods §4-5.

References

1. Lio J, Dong H, Ye Y, Cooper B, Reddy S, Sherer R: **Standardized residency programs in China: perspectives on training quality**. *Int J Med Educ* 2016, **7**:220-221; doi:10.5116/ijme.5780.9b85; PMC4958345.
2. Wu L, Wang Y, Peng X, Song M, Guo X, Nelson H, Wang W: **Development of a medical academic degree system in China**. *Med Educ Online* 2014, **19**:23141; doi:10.3402/meo.v19.23141; PMC3895259.
3. Borakati A: **Evaluation of an international medical E-learning course with natural language processing and machine learning**. *BMC Med Educ* 2021, **21**(1):181; doi:10.1186/s12909-021-02609-8; PMC7992837.
4. Grootendorst M. **BERTopic: neural topic modeling with a class-based TF-IDF procedure**. *arXiv*. 2022; 2203.05794 (accessed 2026-05).
5. Chiu CC, Wu CM, Chien TN, Kao LJ, Li C. **Predicting ICU readmission from electronic health records via BERTopic with long short term memory network approach**. *J Clin Med*. 2024; 13(18):5503.
6. Karabacak M, Jagtiani P, Zipser CM, Tetreault L, Davies B, Margetis K. **Mapping the degenerative cervical myelopathy research landscape: topic modeling of the literature**. *Global Spine J*. 2025; 15(3):1662-1675.
7. Gilardi F, Alizadeh M, Kubli M. **ChatGPT outperforms crowd workers for text-annotation tasks**. *Proc Natl Acad Sci U S A*. 2023; 120(30):e2305016120.
8. Tornberg P. **Best practices for text annotation with large language models**. *arXiv*. 2024; 2402.05129 (accessed 2026-05).
9. Roder M, Both A, Hinneburg A. **Exploring the space of topic coherence measures**. *Proceedings of WSDM*. 2015; 399–408.
10. Dieng AB, Ruiz FJR, Blei DM. **Topic modeling in embedding spaces**. *Trans Assoc Comput Linguist*. 2020; 8:439-453.
11. Pangakis N, Wolken S, Fasching N. **Automated annotation with generative AI requires validation**. *arXiv*. 2023; 2306.00176 (accessed 2026-05).
12. Landis JR, Koch GG. **The measurement of observer agreement for categorical data**. *Biometrics*. 1977; 33(1):159-174.
13. Kuhn HW. **The Hungarian method for the assignment problem**. *Naval Res Logist Q*. 1955;2(1-2):83-97.
14. McInnes L, Healy J, Melville J. **UMAP: uniform manifold approximation and projection for dimension reduction**. *arXiv*. 2018; 1802.03426 (accessed 2026-05).
15. Campello RJGB, Moulavi D, Sander J. **Density-based clustering based on hierarchical density estimates**. *Pacific-Asia Conf Knowl Discov Data Min*. 2013; 160-172.
16. Reimers N, Gurevych I. **Sentence-BERT: sentence embeddings using Siamese BERT-networks**. *arXiv*. 2019; 1908.10084 (accessed 2026-05).
17. Xiao S, Liu Z, Zhang P, Muennighoff N. **C-Pack: packaged resources to advance general Chinese embedding**. *arXiv*. 2023; 2309.07597 (accessed 2026-05).
18. Sun J. **jieba: Chinese text segmentation**. *GitHub*. 2023. Available from: https://github.com/fxsjy/jieba (accessed 2026-05).

19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. **Scikit-learn: machine learning in Python**. *J Mach Learn Res*. 2011; 12:2825–2830.
20. Rehurek R, Sojka P. **Software framework for topic modelling with large corpora**. *Proceedings of LREC*. 2010;45-50.
21. Cohen J. **A coefficient of agreement for nominal scales**. *Educ Psychol Meas*. 1960; 20(1):37-46.
22. Hubert L, Arabie P. **Comparing partitions**. *J Classif*. 1985; 2(1):193-218.
23. Strehl A, Ghosh J. **Cluster ensembles: a knowledge reuse framework for combining multiple partitions**. *J Mach Learn Res*. 2002; 3:583-617.
24. Iniesta R, Stahl D, McGuffin P. **Machine learning, statistical learning and the future of biological research in psychiatry**. *Psychol Med*. 2016; 46(12):2455-2465.
25. Banerjee S, Pedersen T. **The design, implementation, and use of the Ngram Statistics Package**. *Proceedings of CICLing*. 2003; 370-381.
26. Aletras N, Stevenson M. **Evaluating topic coherence using distributional semantics**. *Proceedings of IWCS*. 2013; 13-22.
27. Newman D, Lau JH, Grieser K, Baldwin T. **Automatic evaluation of topic coherence**. *Proceedings of NAACL HLT*. 2010; 100-108.