# Vision-Language Model Guided Semantic Curation for Large-Scale Medical Data Cohorts

Xiaoliang Wang[1](✉), Jiayi Wang[1], Hadrien Reynaud[1], Luca Hagen[1], Ibrahim Ethem Hamamci[2], Sezgin Er[2], Suprosanna Shit[2], Weitong Zhang[3], Bjoern Menze[2], Andreas Maier[1], and Bernhard Kainz[1,3]

[1] FAU Erlangen-Nürnberg, Erlangen, Germany, xiaoliang.wang@fau.de
[2] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland
[3] Department of Computing, Imperial College London, London, UK

**Abstract.** Large-scale datasets are increasingly leveraged for training medical foundation models; however, their anatomical annotations are typically derived from metadata fields that are frequently incomplete, inconsistent, or corrupted. Manual curation of these datasets is prohibitively expensive, and existing single-modality automated methods are brittle when individual data channels fail. Such semantic misalignment between images, reports, and metadata severely limits dataset reliability and undermines downstream model performance. To address this structural bottleneck, we propose a fully automatic, multi-modal reconciliation framework for the semantic curation of large-scale CT repositories. Our approach extracts heterogeneous evidence from multiple sources, including 3D CT volumes, organ presence signals from automated whole-body segmentation, acquisition metadata, and radiology reports. We leverage a Large Language Model (LLM) as an agentic reconciliation module to integrate complementary signals and resolve cross-modal conflicts. This yields globally coherent anatomical region assignments without human supervision. We validate our framework on a large-scale clinical CT repository (>360K patients), demonstrating robust performance even under severe label noise. Our approach substantially outperforms both metadata-driven baselines and state-of-the-art body-part regression models, establishing the necessity of agentic, image-grounded semantic reconciliation for scalable dataset curation. Code is available at https://github.com/xlwang208/VLM-SC-Med/.

**Keywords:** Automatic Dataset Curation · VLM Agents

## 1 Introduction

Medical foundation models rely on massively scaled data cohorts [13,16]. While initiatives like the UK Biobank [14] and CT-RATE [8] drive this frontier, the latter targeting over 300,000 CTs [26], this data explosion suffers from the pervasive presence of intrinsic label noise.

In modern imaging repositories, up to 15% of `BodyPartExamined` DICOM fields are incorrect or missing, posing a critical, systemic threat to downstream

foundation model training [7]. Manual data curation becomes computationally and economically infeasible [22,29]. Historically, pipelines blindly assumed structured metadata accurately reflected the pixel data. Empirical audits repeatedly falsify this: time constraints, institutional mergers, and "protocol drift" render DICOM fields highly unreliable across cohorts [5,3].

Radiology reports, however, are an equally inadequate substitute for a "true" ground truth. While modern Large Language Models (LLMs) can extract clinical entities precisely [9,1,4], clinical reports are inherently constrained to the primary diagnostic indication. Additionally, radiology reports are generated at the study level, while CT studies often contain multiple series with heterogeneous spatial coverage and acquisition parameters. Anatomical regions described in the report may be absent from a given image series, leading to report–image mismatches. Consequently, metadata lies through administrative error, clinical reports are biased by selective description and study-level aggregation, and only the volumetric image provides the most reliable representation of the underlying physical reality. Addressing this tripartite dissonance demands more than naive ensemble classifications or unweighted majority voting schemas, which inevitably fail when administrative artifacts systematically override visual evidence.

To overcome these structural limitations, we propose Agentic Semantic Curation. Inspired by agentic AI [28], our system triangulates visual, textual, and structured metadata signals into a unified reconciliation process. Instead of treating anatomical labeling as a static classification problem, our framework employs an LLM-based reasoning agent that aggregates heterogeneous evidence sources and produces globally consistent anatomical assignments. The agent explicitly summarizes per-modality evidence and justifies relabeling decisions, enabling transparent and fully automatic curation at scale. Our **contributions** are:

1. The first systematic, machine-driven audit of anatomical label reliability across metadata, reports, and image-derived anatomical evidence, quantifying metadata failure as a structured dataset-level problem.
2. A multi-evidence reasoning pipeline that triangulates signals with structured explanations, fundamentally shifting labeling from statistical classification to logical reconciliation.
3. We demonstrate that our framework surfaces systematic miscoding patterns, and exhibits substantially high semantic consistency via a probing-based evaluation on an expert-annotated subset.

**Related Work.** Large-scale imaging archives exhibit systematic label noise. DICOM metadata is often missing or incorrect [7,5], class-conditional noise is common across benchmarks [17], and medical pipelines suffer from correlated annotation errors [10], making manual correction infeasible at the scale of UK Biobank [14] or CT-RATE [8]. Anatomical coverage identification underlies CT analysis. Existing methods include unsupervised regression [27], self-supervised slice ordering [21], the widely adopted DKFZ regressor [20], TotalSegmentator [24], supervised classifiers [19,18], and anatomical positional embeddings [6], yet all rely on single modalities and fail under corrupted or missing signals.

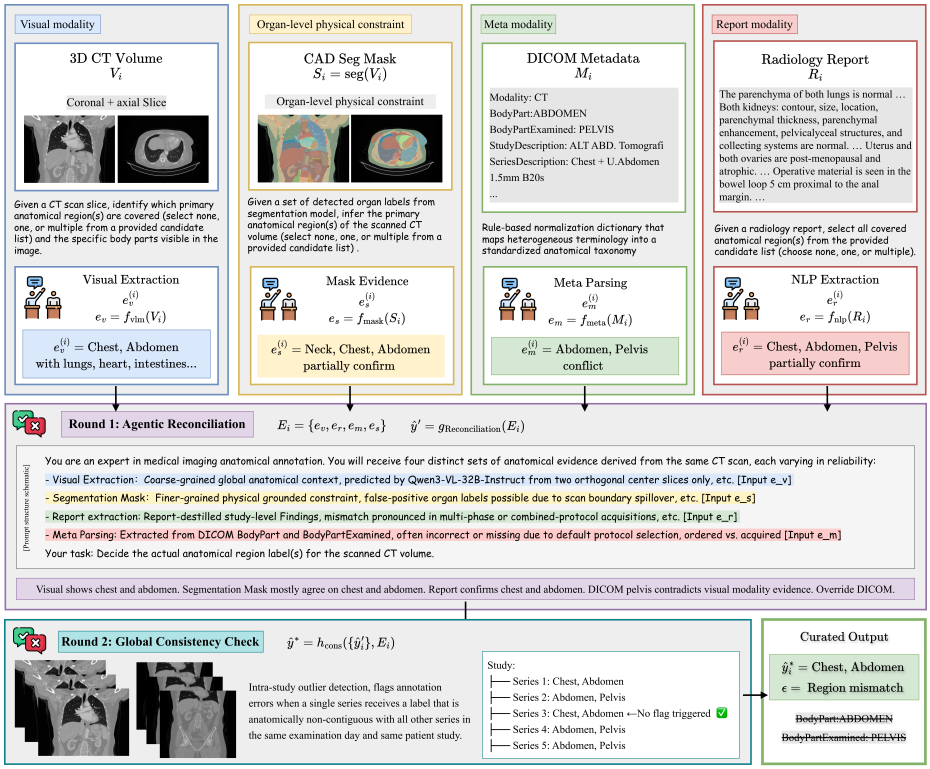


Fig. 1: **Agentic Semantic Curation Pipeline.** Given a study $(V_i, R_i, M_i)$, evidence signals $E_i = \{e_v^{(i)}, e_s^{(i)}, e_m^{(i)}, e_r^{(i)}\}$ are extracted. **Round 1:** an LLM reconciliation module produces a preliminary label $\tilde{y}_i = g_{\text{Reconciliation}}(E_i)$ under organ-level constraints. **Round 2:** a global consistency module yields the final label $\hat{y}_i = h_{\text{cons}}(\tilde{y}_i, E_i)$, and an error taxonomy $\epsilon_i$ recording overridden sources.

Report-based approaches such as CheXpert [9], LEAVS [1], and in-context learning [11], as well as 3D vision-language models including Merlin [2], BioMedCLIP [30], unified multi-source VLMs [12], and RadFM [25], assume clean paired data and lack explicit conflict resolution. We therefore cast label reconciliation as a reasoning problem. Building on ReAct [28], we introduce a constraint semantic decision module that aggregates visual, textual, and metadata evidence to resolve conflicts and produce globally consistent anatomical labels.

## 2 Method

Let $\mathcal{D} = \{(V_i, R_i, M_i)\}_{i=1}^{N}$ denote a heterogeneous imaging corpus of $N$ studies, where each study consists of a 3D CT volume $V_i \in \mathbb{R}^{H \times W \times D}$, an unstructured radiology report $R_i$, and structured DICOM metadata $M_i$. Our objective is to

infer the true anatomical coverage $\hat{y}_i \in \mathcal{Y}$ (e.g., chest, abdomen-pelvis) and to characterize intrinsic label noise without relying on manual annotations.

**Multi-Source Evidence Extraction** For each study $i$, we independently extract anatomical evidence from the visual modality, the segmentation-derived organ space, the radiology report, and the DICOM metadata. This stage converts each modality into structured, comparable signals.

**Visual modality** Performing full 3D multimodal inference on every CT volume is computationally expensive. Instead, each volume $V_i$ is first reoriented into the canonical Right–Anterior–Superior (RAS) coordinate system via affine transformation, and two orthogonal center slices (coronal and axial) are extracted as a lightweight proxy for global anatomical coverage. Following radiological viewing conventions, standard preprocessing is applied, including orientation-based left–right flipping, spacing-aware aspect correction to approximate true anatomical proportions, and isotropic resizing, before concatenating the two slices into a single 2D mosaic passed to the VLM. This step produces visual evidence $e_v^{(i)}$ consisting of a predicted anatomical label and a set of recognized structures. To obtain deterministic physical evidence independent of language interpretation, we apply a pretrained whole-body segmentation model to each CT volume. Specifically, we employ the CADS framework [26], which provides robust multi-organ segmentation across a large anatomical taxonomy. For each volume, we generate a segmentation label map $S_i = \text{seg}(V_i)$ and convert it into a binary organ presence vector $e_s^{(i)} \in \{0,1\}^K$ using a fixed mapping between label IDs and anatomical names, where $K = 167$ corresponds to the number of anatomical categories defined by the CADS framework. This vector reflects which organs are physically present within the scan and serves as a structured physical constraint.

**Metadata modality** We parse raw DICOM headers while filtering out generic protocol-related fields. Particular attention is given to tags that commonly encode anatomical coverage, such as `BodyPart` and `BodyPartExamined`. Because these fields often contain vendor-specific abbreviations or institutional shorthand, we apply a rule-based normalization dictionary that maps heterogeneous terminology into a standardized anatomical taxonomy. For example, "CAP W/O CONT" is normalized to chest-abdomen-pelvis. The normalized anatomical label is represented as metadata evidence $e_m^{(i)}$.

**Report modality** The unstructured radiology report is extracted from the Radiology Information System (RIS). An LLM-based NLP module identifies the anatomical region actually evaluated by the radiologist. Importantly, the model is instructed to analyze the semantic content of the Findings section rather than relying solely on the study title or exam header. This mitigates omission bias in titles. For instance, a report titled "CT Chest" that includes detailed discussion of adrenal glands, kidneys, and hepatic parenchyma indicates chest-abdomen coverage despite the limited title. This step produces report evidence $e_r^{(i)}$, consisting of a structured anatomical label and a list of mentioned organs.

**Agentic Reconciliation** We formulate semantic resolution as an evidence reconciliation problem governed by an LLM agent ($\phi$), which adjudicates the multimodal evidence $\mathbf{E}_i = \{e_v^{(i)}, e_s^{(i)}, e_m^{(i)}, e_r^{(i)}\}$ using clinical priors, as opposed to

naive majority voting, which fails when metadata and reports share incorrect conventions, yielding a validated label and structured reasoning trace: $\tilde{y}_i, \epsilon_i = \arg\max_{y,\epsilon} P_\phi(y, \epsilon \mid \mathbf{E}_i, \mathcal{P})$, where $\tilde{y}_i$ denotes the reconciled anatomical label and $\epsilon_i$ a structured error taxonomy recording modality-specific conflicts and overrides. The prior knowledge $\mathcal{P}$ encodes clinically grounded reliability assumptions across modalities: visual predictions $e_v^{(i)}$ act as a global anatomical context estimator that provides a direct image-level estimate of scan coverage; segmentation-derived organ presence $e_s^{(i)}$ is treated as a physical boundary constraint, providing a finer-grained anatomical constraint by encoding the presence of up to $K = 167$ anatomical structures across the body. Report-distilled labels $e_r^{(i)}$ reflect interpretive study-level semantic evidence. Metadata-derived labels $e_m^{(i)}$ are soft administrative signals that may be incomplete or institution-dependent. When conflicts arise, the agent weighs the reliability of each source. For example, if the report mentions upper abdominal structures and segmentation detects the liver, but metadata indicates "CT Abdomen", the agent recognizes that the scan extends beyond abdomen alone and resolves the label accordingly.

**Global Consistency Enforcement** The final stage operates at the patient-study level to ensure consistency across image series acquired during the same examination date. For this, $\hat{y}_i = h_{\text{cons}}(\tilde{y}_i, \mathbf{E}_i)$ refines the reconciled labels $\tilde{y}_i$ by incorporating hierarchical constraints. A single DICOM Study frequently contains multiple volumetric series with similar anatomical coverage acquired at the same visit (e.g., the Non-contrast, Arterial, Portal venous, and Delayed phases of a Multi-phase liver protocol). We verify that labels assigned to series within the same study are logically consistent: a series whose predicted label is both anatomically non-contiguous with and isolated among all other same-day, same-patient series is flagged as a potential annotation error. For example, an isolated head label within an otherwise consistent abdominal multiphase study is highly suspicious and is examined for potential misclassification. Anatomically adjacent or overlapping region labels are not flagged, *i.e.*, chest-abdomen and abdomen-pelvis within the same study reflect legitimate multi-region acquisitions and are preserved as consistent. Through this three-stage approach, we construct a curated dataset with deterministic anatomical coverage labels, structured multimodal evidence, and traceable reasoning records for every resolved conflict.

## 3 Experiments and Results

**Dataset:** The primary testbed for this evaluation is a large-scale heterogeneous clinical CT repository collected at a local healthcare institution. The dataset is conceptually aligned with the scale and heterogeneity of CT-RATE [8,23], but spans a broader anatomical range from head and neck to pelvis. Studies include both localized scans (e.g., nasal region) and extended field-of-view scans covering partial or complete head-to-pelvis anatomy. The repository contains 360,715 patients collected between 2022 and 2024 across multiple scanner vendors and acquisition protocols, with slice thicknesses ranging from 0.25 mm to 4 mm. The dataset is not publicly available and is used solely for research evaluation.

Table 1: **Evaluation of Anatomical Label Quality on the Expert-Adjudicated Subset.** Class-wise F1 scores for primary anatomical regions alongside global classification metrics evaluated on the expert-adjudicated subset. The proposed Agentic Pipeline achieves the highest overall macro-F1, precision and recall, indicating improved semantic consistency of the curated labels. Best results highlighted ( **bold** ), second best underlined.

| | Class-wise F1 Score | | | | | Global Labeling Metrics | | |
|---|---|---|---|---|---|---|---|---|
| **Curation Strategy** | Head | Neck | Chest | Abdomen | Pelvis | Macro F1 | Precision | Recall |
| Raw Metadata (DICOM) | $0.9091_{\pm 0.0294}$ | $0.3363_{\pm 0.2973}$ | $0.9509_{\pm 0.0135}$ | $0.7095_{\pm 0.0442}$ | $0.0000_{\pm 0.0000}$ | $0.5811_{\pm 0.060}$ | $0.6749_{\pm 0.097}$ | $0.5346_{\pm 0.0518}$ |
| Report-Only (NLP) | $0.9400_{\pm 0.0234}$ | **$0.7726_{\pm 0.1902}$** | $0.9197_{\pm 0.0158}$ | $0.5249_{\pm 0.0371}$ | $0.7195_{\pm 0.0411}$ | $0.7753_{\pm 0.038}$ | $0.7255_{\pm 0.0445}$ | $0.9086_{\pm 0.0316}$ |
| DKFZ Body-Part Reg. | $0.8585_{\pm 0.0351}$ | $0.3605_{\pm 0.1606}$ | $0.9778_{\pm 0.0091}$ | **$0.8638_{\pm 0.028}$** | $0.9083_{\pm 0.0232}$ | $0.7938_{\pm 0.0324}$ | $0.7646_{\pm 0.0254}$ | $0.8855_{\pm 0.0541}$ |
| VLM-Only (Zero-Shot) | $0.9454_{\pm 0.0211}$ | $0.0000_{\pm 0.0000}$ | **$0.9787_{\pm 0.0088}$** | $0.8559_{\pm 0.0284}$ | $0.9067_{\pm 0.0226}$ | $0.7373_{\pm 0.0087}$ | $0.7044_{\pm 0.0134}$ | $0.7772_{\pm 0.0070}$ |
| **Agentic Pipeline** | **$0.9458_{\pm 0.0205}$** | $0.4734_{\pm 0.1964}$ | $0.9717_{\pm 0.0098}$ | $0.8626_{\pm 0.0283}$ | **$0.9092_{\pm 0.0219}$** | **$0.8326_{\pm 0.0398}$** | **$0.7714_{\pm 0.0387}$** | **$0.9334_{\pm 0.0538}$** |

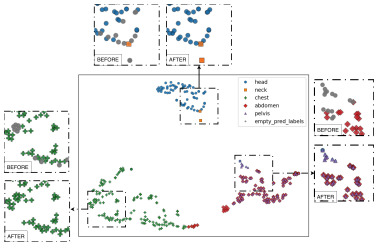


Fig. 2: t-SNE before and after curation. Grey unresolved labels form anatomically coherent clusters post-curation.

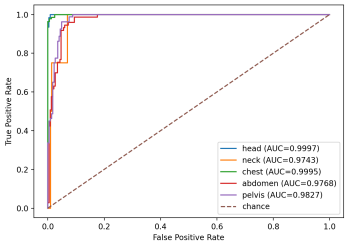


Fig. 3: High AUC across all regions, *i.e.* agreement with expert annotations.

**Implementation Details:** Label generation was performed in a fully unsupervised manner without using expert annotations. All experiments were conducted on a server equipped with 4× NVIDIA A100 80 GB PCIe GPUs. The VLM (Qwen2.5-VL-32B-Instruct, 32 B parameters, bfloat16 precision) was deployed locally using vLLM v0.14.1 with tensor parallelism across all GPUs, following sustained throughputs: report distillation 12.27 samples/s, visual evidence interpretation 9.90 samples/s, and agentic reconciliation 6.49 samples/s. Linear probe classifiers were trained on image embeddings extracted from a pretrained visual encoder (input dimension: 1024), comprising two hidden layers of dimensions 512 and 256 with ReLU activations. Training used the Adam optimizer with learning rate $10^{-3}$ and weight decay $5 \times 10^{-3}$.

**Evaluation:** Four label variants were generated for a 3,000-study subset: **(1)** Raw Metadata labels derived from DICOM `BodyPartExamined`; **(2)** Report-derived labels extracted via LLM-based analysis; **(3)** VLM-only labels obtained from visual projections; **(4)** Agentic-curated labels produced by the proposed reconciliation framework. Embeddings from the 3,000-study subset were paired with automatically generated labels to train an MLP linear probe classifier. The trained probes were then applied to an independent evaluation subset of 375 studies that was manually annotated by board-certified radiologists. Experts re-

viewed the full 3D image volumes to establish gold-standard anatomical labels. These annotations were used exclusively for evaluation. Label quality is measured by probe generalisation: semantically consistent labels produce probes that transfer well to expert annotations, while noisy or semantically inconsistent labels do not. Raw Metadata, Report-Only, and DKFZ Body-Part Regression [20] predictions were additionally compared directly against expert ground truth.

**Results and Problem Characterization.** Table 1 summarizes anatomical labeling performance on the expert-adjudicated subset. The proposed Agentic Semantic Curation pipeline achieves the highest macro-F1 (0.8326), precision (0.7714), and recall (0.9334), outperforming Raw Metadata (0.5811), VLM-only labeling (0.7373), Report-only extraction (0.7753), and DKFZ body-part regression (0.7938). Single-modality methods exhibit characteristic failure modes. Metadata labels collapse on pelvis and show unstable neck performance. Report-derived labels achieve strong recall and the best neck performance but at the cost of reduced precision. DKFZ regression provides strong localization for abdomen and pelvis but degrades in boundary regions such as neck. VLM-only labeling performs well for large anatomical regions but fails completely for neck scans. Fig. 4 further shows that the CADS Sweep, a rule-based organ-presence mapping weighted by organ proportion, hits a performance ceiling at $F1 = 0.805$ directly against expert ground truth, and thereafter only trades recall for precision, confirming that cross-modal reasoning is necessary for robust curation.

Table 2: **Agentic Error Taxonomy & Prevalence.** Prevalence and mechanistic definitions of metadata error categories identified across the clinical cohort.

| Error Category | Mechanistic Definition | Prev. |
|---|---|---|
| *Missing Label* | `BodyPartExamined` field absent or empty, providing no anatomical info. | 26.59% |
| *Region Mismatch* | Extracted visual/textual anatomy contradicts explicit metadata tag. | 53.56% |
| *Multi-Region Truncation* | Volumetric scan covers multiple regions, metadata lists only one. | 6.39% |
| *Over-Specified Label* | Metadata assigns an anatomical region not supported by image evidence. | 0.26% |

**Error Taxonomy Analysis.** The Agentic pipeline reveals that metadata errors are predominantly structural rather than random (Table 2). Region mismatch is the dominant failure mode, accounting for over half of all identified errors, reflecting systematic divergence between institutional naming conventions and actual anatomical coverage. Missing labels (26.59%) reflect incomplete data entry workflows, while multi-region truncation (6.39%) exposes that single-region metadata fields cannot encode extended field-of-view acquisitions.

**Ablation Study.** We perform an ablation study in Tab. 3. Our approach achieves the highest macro-F1 ($0.8326 \pm 0.0398$) and recall ($0.9334 \pm 0.0538$), demonstrating the effectiveness of multi-modal semantic reconciliation. Removing visual evidence yields the largest degradation, showing that image-grounded constraints are vital for anchoring the agent. Without segmentation-derived evidence, the agent becomes more susceptible to inconsistencies between visual pre-

Table 3: **Ablation Study.** Impact of individual evidence modalities on curation performance; each row removes one input source from the full pipeline. Best results highlighted (**bold**), second best underlined.

| Configuration | F1 Score | Precision | Recall |
|---|---|---|---|
| **Full Pipeline** | **0.8326**$_{\pm 0.03}$ | 0.7714$_{\pm 0.03}$ | **0.9334**$_{\pm 0.05}$ |
| w/o Visual Evid. | 0.7490$_{\pm 0.04}$ | **0.8024**$_{\pm 0.04}$ | 0.8229$_{\pm 0.03}$ |
| w/o Report (NLP) | 0.8193$_{\pm 0.04}$ | 0.7865$_{\pm 0.05}$ | 0.8831$_{\pm 0.05}$ |
| w/o CADS Seg. | 0.7875$_{\pm 0.05}$ | 0.7753$_{\pm 0.07}$ | 0.8309$_{\pm 0.05}$ |

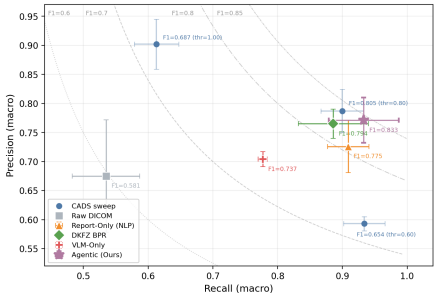


Fig. 4: Precision-Recall Space. Trade-off across curation strategies.

dictions and semantic sources. Removing the free-text report drops performance in complex boundary cases where generic metadata conflicts with visual ambiguity. In contrast, removing report-derived evidence produces only a moderate reduction in macro-F1 (0.8193), indicating that report signals provide complementary but less essential information compared to image-derived evidence.

**Discussion:** Our results show that **Agentic Semantic Curation** reduces structural label noise, outperforming single-modality Raw Metadata (0.5811) and Report-Only (0.7753) baselines with 0.8326 Macro F1 by casting curation as a multi-source reasoning task, confirming that anatomical label noise is fundamentally a *conflict resolution* problem: no individual signal is consistently reliable, and the agent's value lies in adjudicating when sources disagree. Visual grounding resolves omission bias and guards against protocol drift, while a global consistency module enforces longitudinal coherence. Neck classification remains the weakest region across all methods, partly attributable to severe class imbalance: neck-specific acquisitions account for only 1.34–1.97% of studies in our cohort both before and after curation, leaving the linear probe with limited training signal for this class.

## 4 Conclusion

The integrity of large-scale medical imaging datasets is increasingly compromised by label noise from unreliable metadata, protocol drift, and incomplete reports, a bottleneck that intensifies as foundation models rely on massive uncurated corpora. Our results show that **Agentic Semantic Curation** mitigates these risks by reframing curation as an interpretable reasoning task. By fusing visual, textual, and structured evidence through an autonomous engine, the approach achieves a Macro F1 of 0.8326, outperforming raw metadata and state-of-the-art body-part regression models. This triangulation resolves omission bias by grounding sparse reports in 3D projections and segmentation signals, while providing traceable interpretability through structured error taxonomies. As medical AI scales to million-scan cohorts, this image-grounded reconciliation

framework, a prerequisite enabling more granular downstream tasks such as series description normalization and contrast phase identification, offers a scalable blueprint for high-fidelity data integrity in clinical deployment.

**Acknowledgments.** We acknowledge HPC resources from NHR@FAU (projects b143dc, b180dc), funded by federal and Bavarian state authorities and Gerhard Wellein and his team's HPC approach. NHR@FAU hardware is partially funded by DFG 440719683. Additional support was received from ERC projects MIA-NORMAL 101083647 and CHARMS 101246053, DFG 513220538 and 512819079, and the state of Bavaria (HTA and the Bavarian Foundation Model Initiative). We further acknowledge resources provided by the Isambard-AI National AI Research Resource (AIRR), operated by the University of Bristol and funded by DSIT via UKRI and STFC [ST/AIRR/I-A-I/1023] [15]. We thank the Helmut Horten Foundation for generous support, which made this work possible, and the Istanbul Medipol University for invaluable support and provision of data. This research was supported in part by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health (NIH), and used the computational resources of the NIH high-performance computing Biowulf cluster. We used coding agents and LLMs from Anthropic, OpenAI, Google for coding, experiment orchestration, cluster monitoring, and as a writing aid.

**Disclosure of Interests.** The authors have no relevant competing interests.

## References

1. Bigolin Lanfredi, R., Zhuang, Y., Finkelstein, M., Thoppey Srinivasan Balamuralikrishna, P., Krembs, L., Khoury, B., Reddy, A., Mukherjee, P., Rofsky, N.M., Summers, R.M.: LEAVS: An LLM-based labeler for abdominal CT supervision. In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. LNCS, vol. 15964, pp. 316–326. Springer (2025)
2. Blankemeier, L., Kumar, A., Cohen, J.P., Liu, J., Liu, L., Van Veen, D., Gardezi, S.J.S., Yu, H., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., et al.: Merlin: a computed tomography vision-language foundation model and dataset. Nature **652**, 1318–1328 (2026). https://doi.org/10.1038/s41586-026-10181-8
3. Brzus, M., Riley, C.J., Bruss, J., Boes, A., Jones, R., Johnson, H.J.: DICOM sequence selection for medical imaging applications. In: Proc. SPIE Medical Imaging. vol. 12931, p. 1293108. SPIE (2024). https://doi.org/10.1117/12.3006568
4. Das, A., Talati, I.A., Chaves, J.M.Z., Rubin, D., Banerjee, I.: Weakly supervised language models for automated extraction of critical findings from radiology reports. npj Digital Medicine **8**, 257 (2025). https://doi.org/10.1038/s41746-025-01522-4
5. Gauriau, R., Bridge, C., Chen, L., Kitamura, F., Tenenholtz, N.A., Kirsch, J.E., Andriole, K.P., Michalski, M.H., Bizzo, B.C.: Using DICOM metadata for radiological image series categorization: a feasibility study on large clinical brain MRI datasets. Journal of Digital Imaging **33**, 747–762 (2020)
6. Goncharov, M., Samokhin, V., Soboleva, E., Sokolov, R., Shirokikh, B., Belyaev, M., Kurmukov, A., Oseledets, I.: Anatomical positional embeddings. In: MICCAI 2024. vol. LNCS 15010, pp. 68–77. Springer Nature Switzerland (2024)
7. Gueld, M.O., Kohnen, M., Keysers, D., Schubert, H., Wein, B.B., Bredno, J., Lehmann, T.M.: Quality of DICOM header information for image categorization. In: Proc. SPIE Medical Imaging. vol. 4685, pp. 280–287. SPIE (2002)

8. Hamamci, I.E., Er, S., Wang, C., et al.: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nature Biomedical Engineering (2026). https://doi.org/10.1038/s41551-025-01599-y
9. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019). https://doi.org/10.1609/aaai.v33i01.3301590
10. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis **65**, 101759 (2020). https://doi.org/10.1016/j.media.2020.101759
11. Kim, S., Kim, D., Kim, J., Koo, J., Yoon, J., Yoon, D.: In-context learning with large language models: A simple and effective approach to improve radiology report labeling. Healthcare Informatics Research **31**(3), 295–309 (2025)
12. Lee, H.C., Liu, Z., Ahmed, H., Kim, S., Huver, S., Nath, V., Fayad, Z.A., Deyer, T., Mei, X.: Unified supervision for vision-language modeling in 3D computed tomography. In: Proc. ICCV Workshops. pp. 6784–6792 (2025)
13. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
14. Littlejohns, T.J., Holliday, J., Gibson, L.M., et al.: The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nature Communications **11**, 2624 (2020). https://doi.org/10.1038/s41467-020-15948-9
15. McIntosh-Smith, S., Alam, S.R., Woods, C.: Isambard-AI: a leadership class supercomputer optimised specifically for artificial intelligence. arXiv.2410.11199 (2024)
16. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**, 259–265 (2023). https://doi.org/10.1038/s41586-023-05881-4
17. Northcutt, C.G., Jiang, L., Chuang, I.L.: Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research **70**, 1373–1411 (2021). https://doi.org/10.1613/jair.1.12125
18. Ouyang, Z., Zhang, P., Pan, W., Li, Q.: Deep learning-based body part recognition algorithm for three-dimensional medical images. Medical Physics **49**(5), 3067–3079 (2022). https://doi.org/10.1002/mp.15536
19. Raffy, P., Pambrun, J.F., Kumar, A., Dubois, D., Patti, J.W., Cairns, R.A., Young, R.: Deep learning body region classification of MRI and CT examinations. Journal of Digital Imaging **36**(4), 1291–1301 (2023)
20. Schuhegger, S.: Body part regression for CT images. arXiv preprint arXiv:2110.09148 (2021), master's thesis, German Cancer Research Center (DKFZ)
21. Tang, Y., Gao, R., Han, S., Chen, Y., Gao, D., Nath, V., Bermudez, C., Savona, M.R., Bao, S., Lyu, I., Huo, Y., Landman, B.A.: Body part regression with self-supervision. IEEE Transactions on Medical Imaging **40**(5), 1499–1507 (2021)
22. Thiriveedhi, V.K., Krishnaswamy, D., Clunie, D., Pieper, S., Kikinis, R., Fedorov, A.: Cloud-based large-scale curation of medical imaging data using AI segmentation. Research Square (2024). https://doi.org/10.21203/rs.3.rs-4351526/v1
23. VLM3D Organizers: VLM3D challenge: Vision-language modeling in 3D medical imaging. https://research.forithmus.com/collections/vlm3d-challenge (2025), iCCV 2025 Workshop

24. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024
25. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D & 3D medical data. Nature Communications **16**, 7866 (2025)
26. Xu, M., Amiranashvili, T., Navarro, F., et al.: CADS: A comprehensive anatomical dataset and segmentation for whole-body anatomy in computed tomography. arXiv:2507.22953 (2025)
27. Yan, K., Lu, L., Summers, R.M.: Unsupervised body part regression via spatially self-ordering convolutional neural networks. In: Proc. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1022–1025. IEEE (2018)
28. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: ReAct: Synergizing reasoning and acting in language models. In: The Eleventh International Conference on Learning Representations (2023)
29. Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Jiang, Z., Zhong, S., Hu, X.: Data-centric artificial intelligence: A survey. ACM Computing Surveys **57**(5), 1–42 (2025). https://doi.org/10.1145/3711118
30. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2025). https://doi.org/10.1056/AIoa2400640