излагать свои рассуждения перед выдачей ответа (например, *Chain-of-Thought*) или фреймворки реализующие строгий цикл «рассуждение-действие» (например, *ReAct*) [3]. Это позволяет отслеживать логику и снижает вероятность спонтанных ошибок.

Второй уровень – контроль над действиями. Данные методы верифицируют команду агента перед ее выполнением. Наиболее распространенным подходом является автоматическая валидация сгенерированных вызовов *API* на соответствие заранее определенной схеме (*JSON Schema*, *Pydantic*), которая отбраковывает все синтаксически неверные команды. Для критически важных операций применяется подтверждение со стороны человека (*Human-in-the-Loop*), где агент лишь формирует команду, а ее исполнение требует явного одобрения.

Третий уровень – контроль над последствиями. Эти методы ограничивают потенциальный ущерб от действий агента, даже если они прошли предыдущие проверки. Ключевой техникой является выполнение команд в изолированной среде («песочнице», например, *Docker*-контейнере) с ограниченными правами доступа к сети и файловой системе, а также установка жестких лимитов на количество вызовов *API* или расходуемый бюджет.

Внедрение «*Guardrails*» – это поиск компромисса между двумя ключевыми свойствами агента: автономией (способностью решать сложные задачи) и «выравниванием» (соответствием действий целям и ограничениям человека).

В ходе работы была проанализирована проблема контролируемости автономных *LLM*-агентов и систематизированы основные риски, связанные с их стохастической природой. В качестве результата предложена трехуровневая модель архитектурных паттернов контроля «*Guardrails*», включающая контроль на уровне рассуждений, действий и последствий. Анализ показал, что путь к созданию мощных и безопасных *LLM*-агентов лежит не столько в увеличении размера моделей, сколько в разработке продуманных, многоуровневых архитектур контроля. Дальнейшие исследования должны быть сосредоточены на создании методик, упрощающих проектирование и внедрение таких «*Guardrails*» для безопасного применения *LLM*-агентов в промышленных приложениях.

Литература

1. Курочка, К. С. Нейросетевая модель автогенерации тестов для студентов в системе Moodle на основе анализа методических материалов / К. С. Курочка, Ю. С. Башаримов // Цифровая трансформация. – 2025. – N 31 (3). – P. 66–75. – URL: https://doi.org/10.35596/1729-7648-2025-31-3-66-75 (дата обращения: 11.10.2025).
2. Amodei D., Olah C., Steinhardt J., Christiano P., Schulman J., Mané D. Concrete Problems in AI Safety // ArXiv preprint arXiv:1606.06565. – 2016.
3. Synergizing Reasoning and Acting in Language Models / S. Yao, J. Zhao, D. Yu [et al.] // International Conference on Learning Representations (ICLR), 2023.

METHODOLOGY FOR ASSESSING THE ACCURACY AND QUALITY OF LABELING AND SEGMENTATION OF MRI IMAGES OF THE HUMAN LUMBAR SPINE

Ren Huanhai[1, 2], Wang Xuemei[1, 2], K. S. Kurachka[1]

[1]Sukhoi State Technical University of Gomel, Republic of Belarus

[2]Shandong Huayu Institute of Technology, Dezhou, People's Republic of China

This study proposes a unified evaluation framework for automated lumbar spine MRI segmentation and level-wise labeling using public multi-center data [1]. The protocol assesses detection/segmentation via mAP across IoU thresholds (0.50-0.95), mask quality via Dice and 95 % Hausdorff distance, and labeling accuracy on ROIs. It addresses cross-domain variations and

recommends engineering strategies like a two-stage cascade. This structured approach ensures reproducibility and offers a stratified benchmark for future research.

Keywords: lumbar spine MRI, automated segmentation, level wise labelling, unified evaluation, cross domain generalization.

МЕТОДОЛОГИЯ ОЦЕНКИ ТОЧНОСТИ И КАЧЕСТВА МАРКИРОВКИ И СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ МРТ ПОЯСНИЧНОГО ОТДЕЛА ПОЗВОНОЧНИКА ЧЕЛОВЕКА

Рен Хуанхай[1, 2], Ванг Хуамей[1, 2], К. С. Курочка[1]

[1]Гомельский государственный технический университет имени П. О. Сухого, Республика Беларусь

[2]Шаньдунский технологический университет Хуаюй, г. Дэчжоу, Китайская Народная Республика

В данном исследовании предложена унифицированная система оценки для автоматизированной сегментации и поуровневой маркировки МРТ поясничного отдела позвоночника с использованием общедоступных данных [1]. Предлагаемая методика позволяет оценить локализацию/сегментацию с помощью mAP при пороговых значениях IoU (0,50–0,95), качество маски по методу Дайса и 95%-ному расстоянию Хаусдорфа, а также точность маркировки в областях интереса. В ней рассмотрены междоменные вариации и рекомендованы инженерные стратегии, такие как двухступенчатый каскад. Этот структурированный подход обеспечивает воспроизводимость и предлагает стратифицированный ориентир для будущих исследований.

Ключевые слова: МРТ поясничного отдела позвоночника, автоматизированная сегментация, маркировка по уровням, унифицированная оценка, междоменное обобщение.

Deep learning for lumbar spine MRI analysis localizes, segments, and labels vertebrae and discs to reduce manual workload and improve consistency [2]. While multistage pipelines show promise, robustness across imaging centers and sequences remains challenging. Public multi-center benchmarks, such as a dataset of 447 T1/T2 sagittal series, provide reference segmentations and labels, enabling cross-domain evaluation. A unified evaluation protocol is adopted: mAP over IoU thresholds from 0.50–0.95 for detection and segmentation, combined Dice and 95% Hausdorff distance for mask quality, and region-level accuracy for labeling. Following nnU-Net practices, standardized preprocessing and training are implemented to ensure reproducibility and fair comparison.

1. Public Resources and Evaluation Benchmarks

The SPIDER [1] benchmark serves as the principal public resource for reproducible research in lumbar spine MRI analysis. It comprises 447 T1/T2 sagittal series from 218 patients across four hospitals, providing reference segmentations for vertebral bodies, intervertebral discs, and the spinal canal, along with level-wise annotations. The dataset includes fixed training/validation splits and a held-out test set to facilitate fair cross-vendor and cross-protocol comparisons. Additionally, a smaller multi-scanner dataset of 34 cases is available for assessing cross-domain generalization. SPIDER is integrated with the Grand Challenge platform, enabling online evaluation while keeping test labels confidential.

We adopt a standardized labeling taxonomy: vertebrae L1–L5 and discs from L1–L2 to L5–S1 [2, 3]. Ambiguity from lumbosacral transitional vertebrae is resolved using the iliolumbar ligament as the anatomical landmark. To mitigate confounding effects from pathology-induced boundary ambiguity, we emphasize precise boundary definitions and recommend stratified performance reporting by sequence and scanner vendor.

2. Unified Evaluation Protocol

This study uses a consistent protocol: detection/instance segmentation use mAP (0.50-0.95 IoU, plus 0.50/0.75 and per-class results); mask quality uses Dice and 95th-percentile Hausdorff distance (by anatomy/level); level-wise labeling uses Top-1 accuracy/confusion matrix (vertebrae/discs separately). It follows COCO for detection.

Protocol details: patient-level splits (with IDs/seeds), specified preprocessing/augmentation, class imbalance solutions (report per-class/overall results), domain shift quantification (by vendor/T1-T2), and released metric scripts/model info for reproducibility.

3. Cross-domain generalization: sources and strategies

In multi-center lumbar spine MRI, domain shift stems from acquisition (vendor, T1/T2, resolution, etc.) and annotation (nomenclature/boundary inconsistencies). Results should be stratified by vendor/sequence/resolution, with harmonized terminology.

Mitigation: two-stage cascade (high-recall detector + high-fidelity segmenter), topology-aware postprocessing, class resampling/loss weighting (rare levels), test-time augmentation. These boost cross-domain stability under unified reporting.

This study proposes a unified lumbar spine MRI evaluation framework using public multi-center data, employing mAP (IoU 0,50–0,95) for detection/segmentation, Dice and 95 % Hausdorff distance for mask quality, and level accuracy within ROIs to quantify domain shifts and enhance reproducibility. Future work should implement patient-level splits, stratified reporting by vendor/sequence, and a two-stage cascade with topology-constrained postprocessing, augmented by resampling for rare levels and test-time augmentation. Subsequent studies should target per-class mAP $\geq$ 0,80 and level accuracy $\geq$ 97 % using representative models to ensure clinically applicable results.

References

1. SPIDER – URL: https://zenodo.org/records/10159290, VB/IVD DB: https://osf.io/-qx5rt/ (date of access: 11.10.2025).
2. Kanstantsin Kurachka, Ren Huanhai. Intelligent System for Analyzing MRI Images to Find the Main Elements of the Human Spine // 2025 Open Semantic Technologies for Intelligent Systems (OSTIS-2025). – C. 361–366.
3. Kurachka Kanstantsin, Ren Huanhai, Wang Xuemei .Comparative Analysis of Deep Learning Models for Lumbar Vertebrae Segmentation in MRI Image // 2025 Pattern Recongnition And Information Processing (PRIP 2025). – C. 176–179.

RESEARCH ON THE METHOD OF HUMAN LUMBAR SPINE MODEL RECONSTRUCTION BASED ON PUBLIC DATASETS

Wang Xuemei[1, 2], Ren Huanhai[1, 2], K. S. Kurachka[1]

[1]Sukhoi State Technical University of Gomel, Republic of Belarus

[2]Shandong Huayu Institute of Technology, Dezhou, People's Republic of China

This paper aims to describe a standardized reconstruction process for computational biomechanical models of the human lumbar spine based on publicly available CT datasets. The process begins with the geometric and topological information of the publicly available datasets and proceeds through three sequential stages: data preprocessing, geometric model reconstruction, and biomechanical modeling and simulation, ultimately generating a patient-specific model suitable for simulation analysis. This paper systematically discusses the fundamental role of publicly available datasets as data sources in the reconstruction process, and proposes a series of methods for reconstructing computational models of the human lumbar spine. This work provides a clear technical path for the reproducible construction of computational lumbar spine models.