AImageLab

September 9, 2026

AI4M3D · ECCV 2026

ECCV 2026 | MALMÖ | SEPT 8–13

From Volumes to Clinical Decision

Five conditions to fill the gap between medical 3D vision and the clinic

Shared Data · Reliable Supervision · Robust Models

Actionable Outputs · Trusted Evaluation

Federico Bolelli

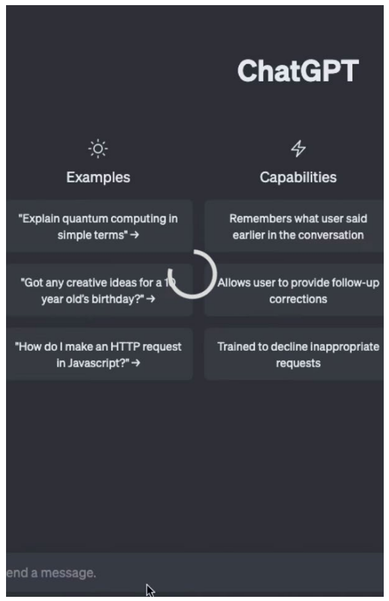
AImageLab — University of Modena and Reggio Emilia · federico.bolelli@unimore.it

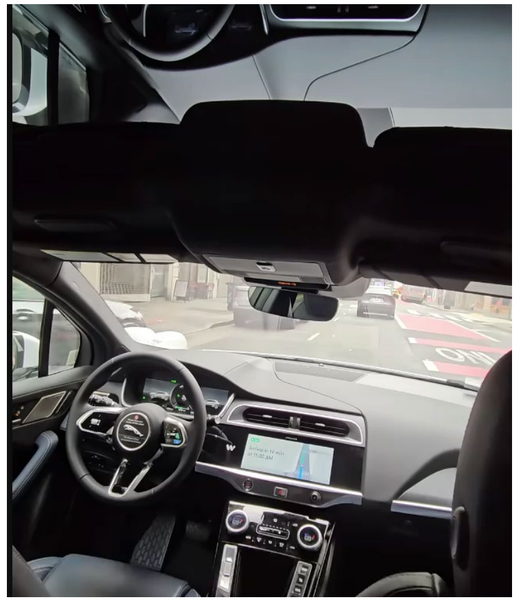
UNIMORE

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImageLab

AI Is Everywhere. Almost.

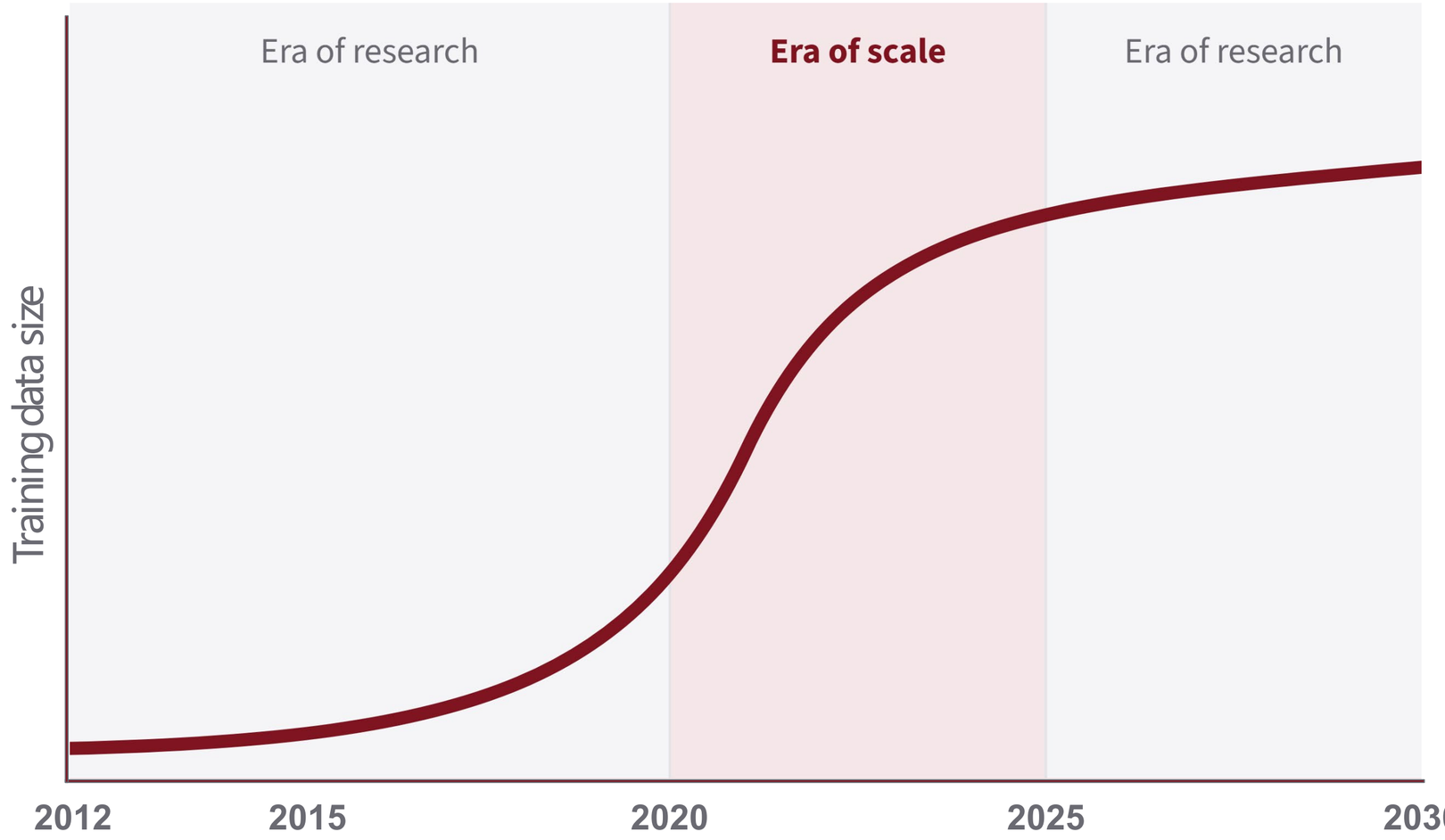




Video not available in the PDF!

Why Medical 3D Vision Has No ImageNet Moment (Yet)?

General models crossed into the era of scale.



VISION–LANGUAGE **650×**

SigLIP 2 **10B**

BiomedCLIP **15.3M**

SEGMENTATION **700×**

SAM **1.1B**

MedSAM **1.57M**

CLASSIFICATION **11×**

ImageNet-21k **14.2M**

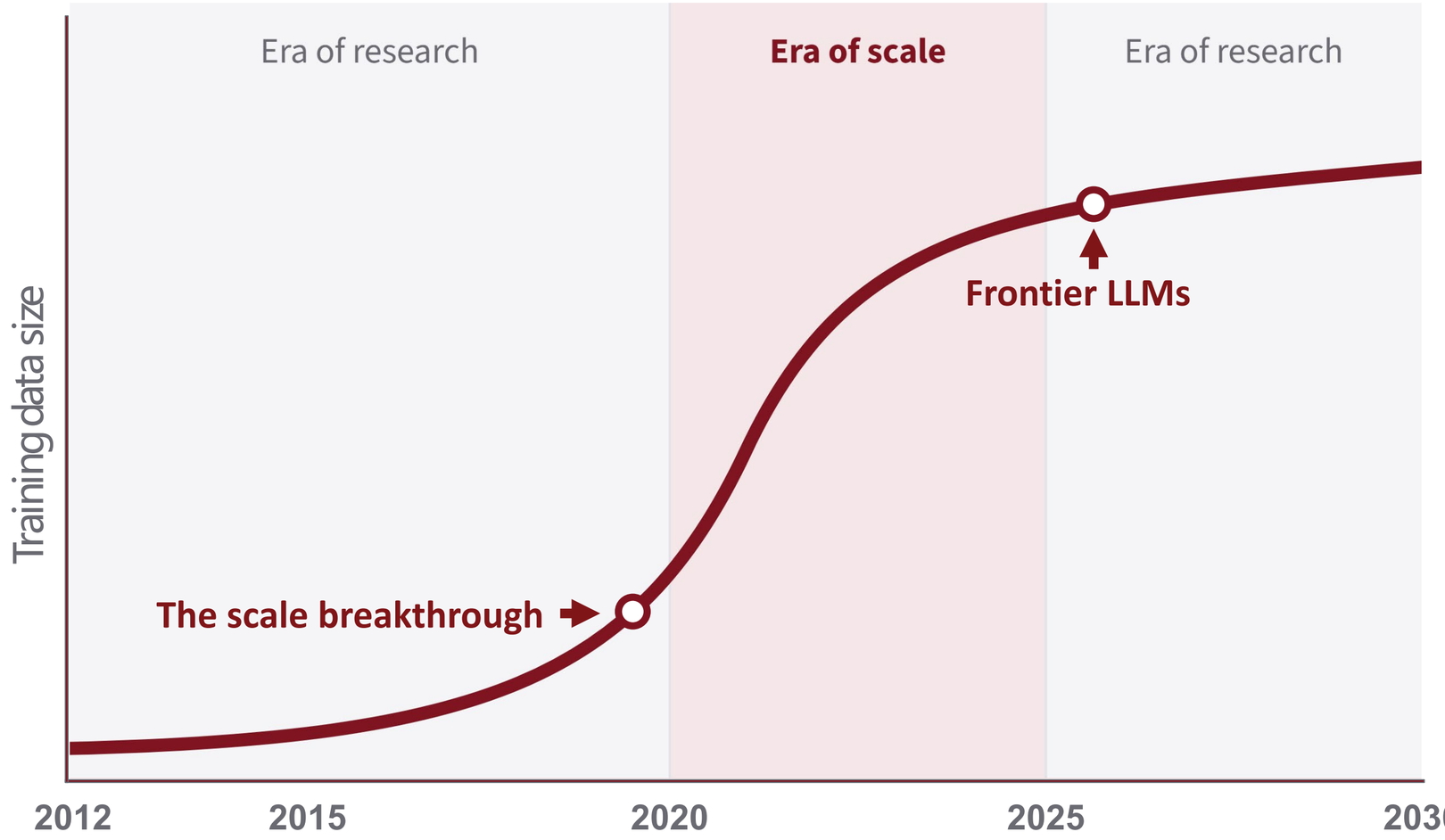
RadImageNet **1.35M**

Counts as reported by the authors: WebLI (SigLIP 2), PMC-15M (BiomedCLIP), SA-1B (SAM), MedSAM, ImageNet-21k, RadImageNet.

[1] Curtis P. Langlotz - Toward a Radiology Foundation Model

Why Medical 3D Vision Has No ImageNet Moment (Yet)?

General models crossed into the era of scale.



VISION–LANGUAGE **650×**

SigLIP 2 10B

BiomedCLIP 15.3M

SEGMENTATION **700×**

SAM 1.1B

MedSAM 1.57M

CLASSIFICATION **11×**

ImageNet-21k 14.2M

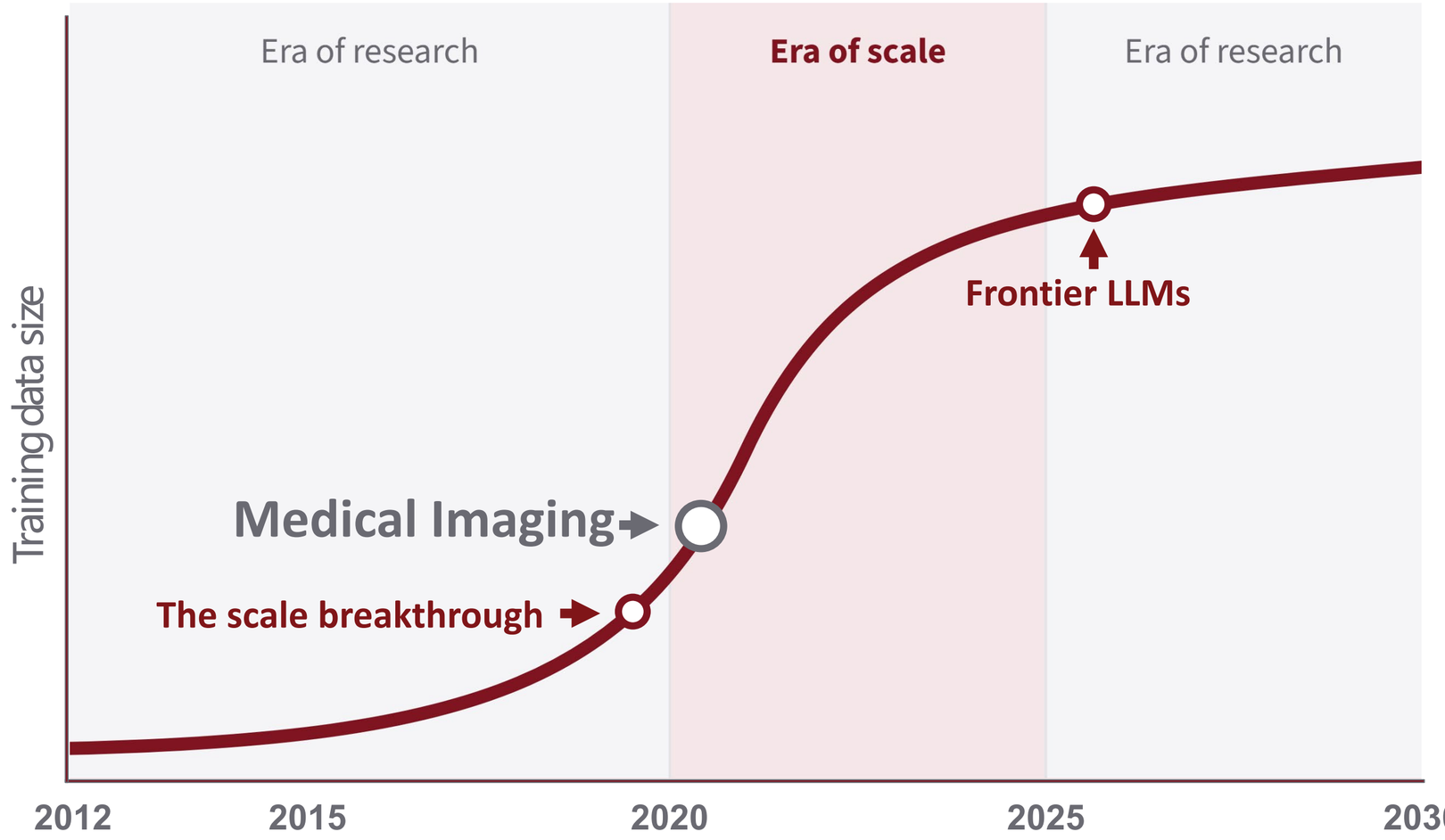
RadImageNet 1.35M

Counts as reported by the authors: WebLI (SigLIP 2), PMC-15M (BiomedCLIP), SA-1B (SAM), MedSAM, ImageNet-21k, RadImageNet.

[1] Curtis P. Langlotz - Toward a Radiology Foundation Model

Why Medical 3D Vision Has No ImageNet Moment (Yet)?

General models crossed into the era of scale.



VISION–LANGUAGE **650×**

SigLIP 2 10B

BiomedCLIP 15.3M

SEGMENTATION **700×**

SAM 1.1B

MedSAM 1.57M

CLASSIFICATION **11×**

ImageNet-21k 14.2M

RadImageNet 1.35M

Counts as reported by the authors: WebLI (SigLIP 2), PMC-15M (BiomedCLIP), SA-1B (SAM), MedSAM, ImageNet-21k, RadImageNet.

[1] Curtis P. Langlotz - Toward a Radiology Foundation Model.

AI as a Second Reader: The Evidence

WHERE IT ALREADY WORKS

PATHOLOGY — BREAST, WSI

0.994 algorithm

0.884 pathologist

AUC of the best algorithm against the best pathologist in a time-constrained diagnostic setting. Seven of 32 algorithms exceeded a panel of 11 pathologists.

DERMOSCOPY — MELANOMA

0.86 CNN

0.79 dermatologists

ROC AUC of a CNN against the mean of 58 dermatologists in a reader study.

BREAST HISTOLOGY

99% panel of four

85% single observer

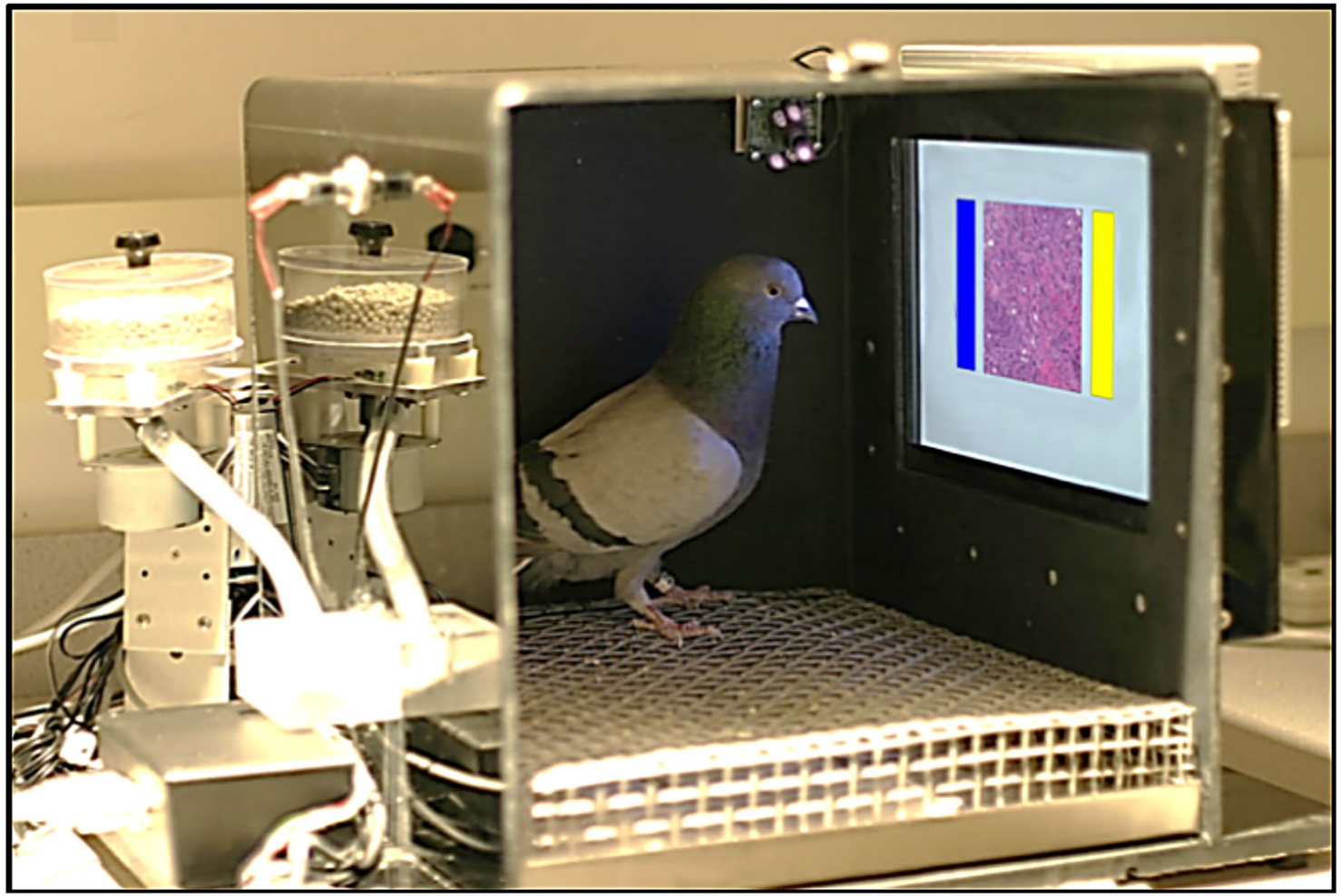
Accuracy on benign versus malignant discrimination on previously unseen slides; pooling four independent observers raises it to 99%.

[1] Ehteshami Bejnordi, Babak, et al. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer" *Jama 2017.*

[2] Haenssle, Holger A., et al. "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists." *Annals of Oncology 2018.*

[3] *Levenson et al.,* "Pigeons (*Columba livia*) as Trainable Observers of Pathology and Radiology Breast Cancer Images" *PLoS ONE 2015.*

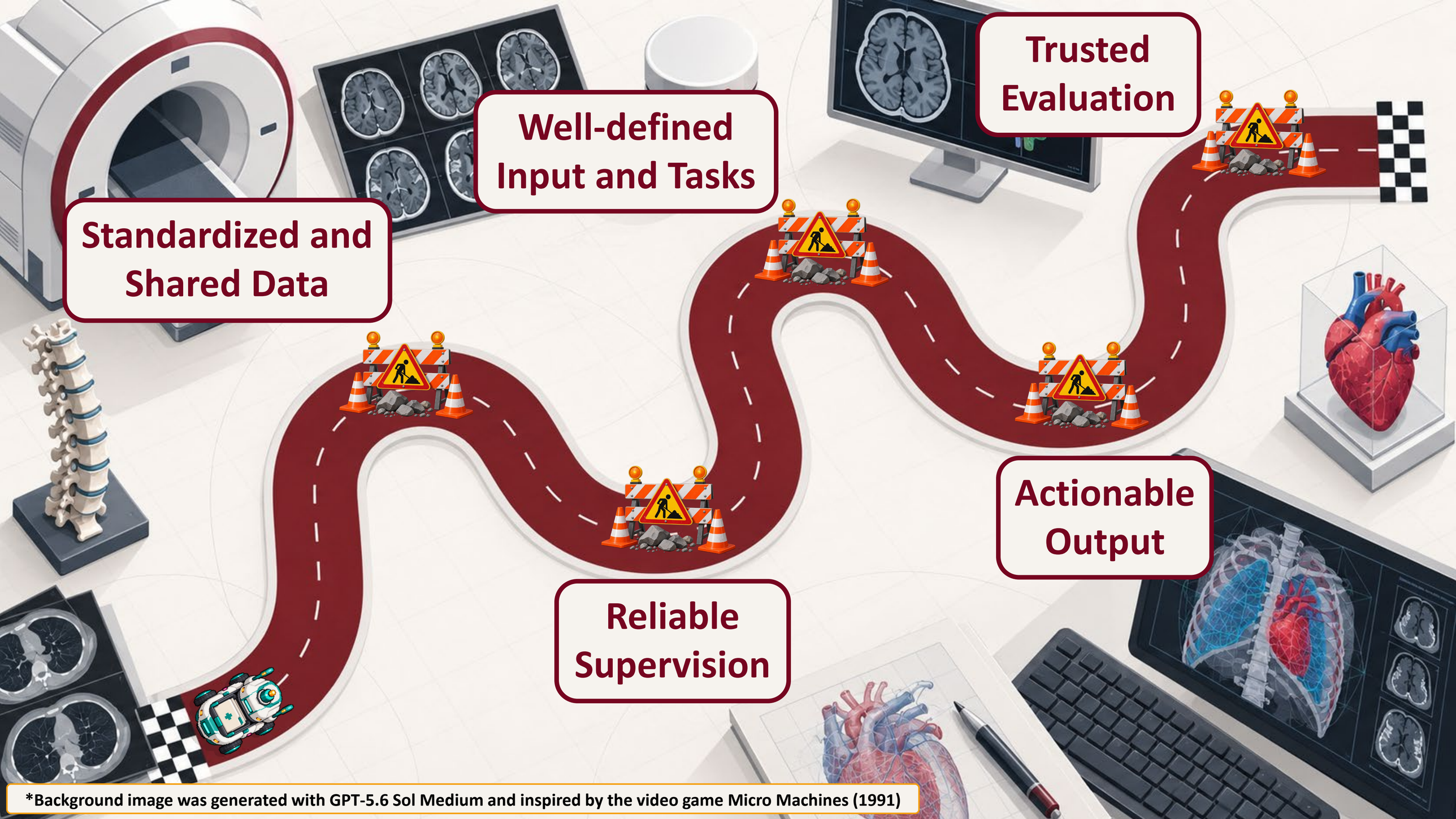
AI as a Second Reader: The Evidence



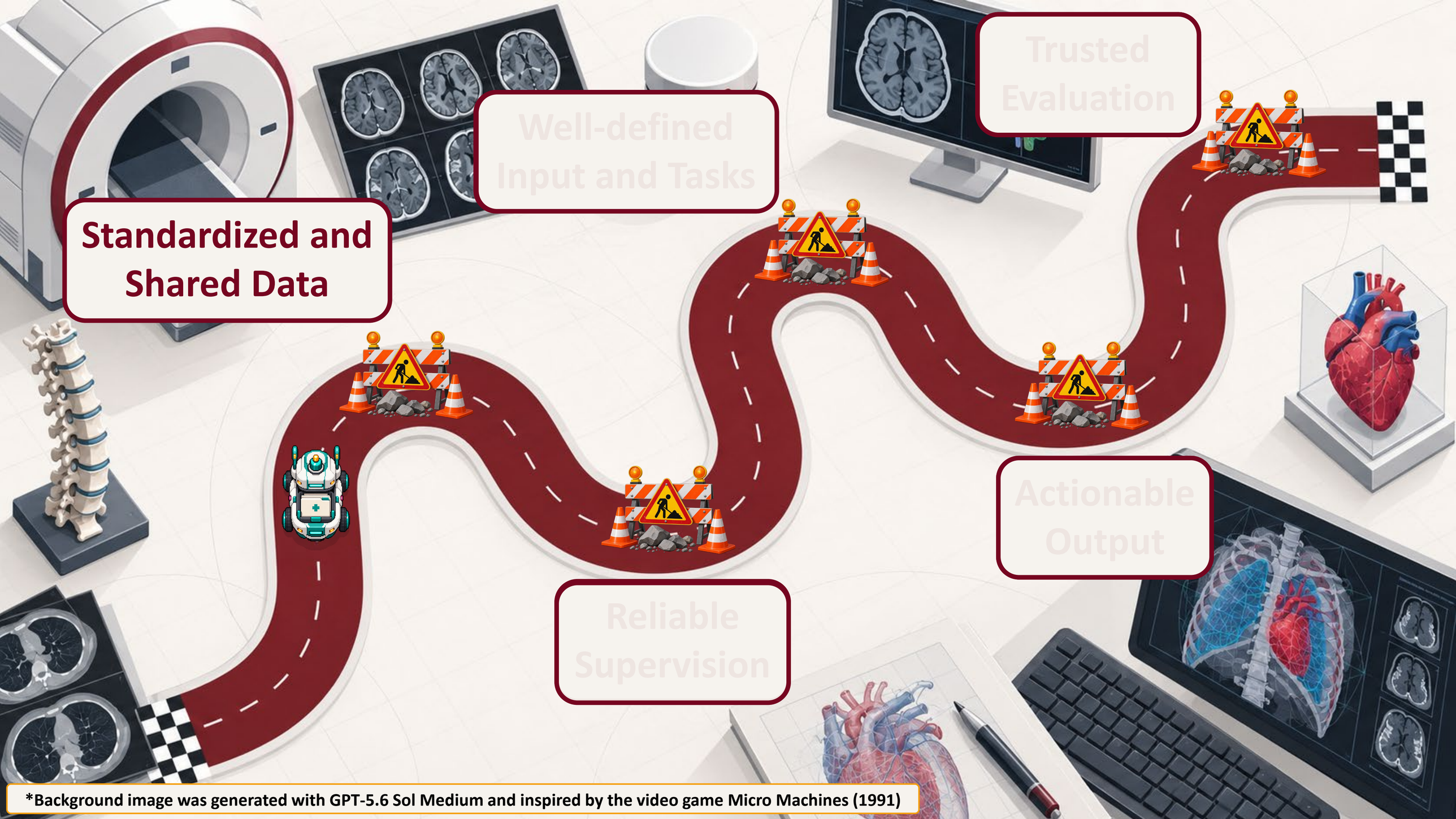
[1] Ehteshami Bejnordi, Babak, et al. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer" *Jama 2017.*

[2] Haenssle, Holger A., et al. "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists." *Annals of Oncology 2018.*

[3] *Levenson et al.,* "Pigeons (*Columba livia*) as Trainable Observers of Pathology and Radiology Breast Cancer Images" *PLoS ONE 2015.*



*Background image was generated with GPT-5.6 Sol Medium and inspired by the video game Micro Machines (1991)



*Background image was generated with GPT-5.6 Sol Medium and inspired by the video game Micro Machines (1991)

Benchmark Standardization

MEDICAL IMAGING BENCHMARKS

Three layers now make methods more comparable.

01

Shared task

Same cohort, image inputs, reference labels and train/test split.

BraTS · CAMELYON · Decathlon

02

Controlled evaluation

Hidden test labels, central scoring and a fixed submission protocol.

MICCAI challenge ecosystem

03

Standardized evidence

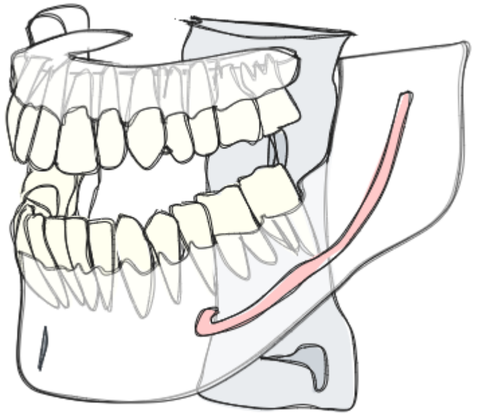
Task-aware metric selection, ranking rules and transparent reporting.

BIAS · Metrics Reloaded

[1] L. Maier-Hein, et al. “Why rankings of biomedical image analysis competitions should be interpreted with care.” Nature Communications 2018.
[2] L. Maier-Hein, et al. “Metrics reloaded: recommendations for image analysis validation.” Nature Methods 2024.

TOOTH FAIRY

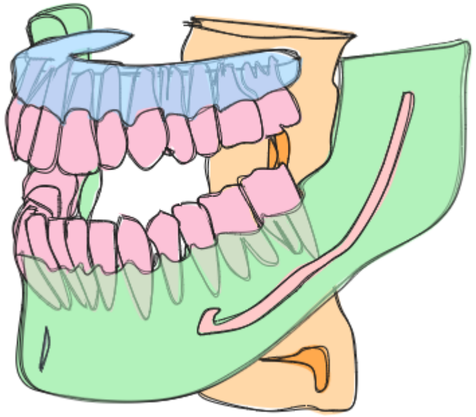
1 class



2023

TOOTH FAIRY2

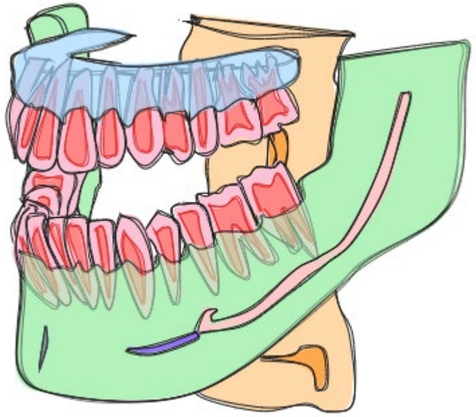
42 classes



2024

TOOTH FAIRY3

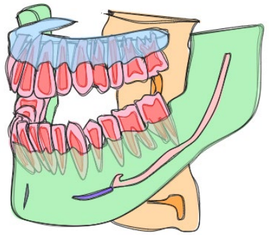
77 classes



2025

TOOTH FAIRY4

77 classes + Reports



REPORT

Complete mandibular dentition; 18 and 28 absent in the maxilla. 36 and 46 likely restored with post-and-core and covered by prosthetic crowns. 15 with endodontic material 2 mm from the apex and an area of radiolucency involving the periradicular alveolar bone; 16 likely restored with post-and-core and covered by a prosthetic crown. Black class I dental fillings of teeth 17, 16, 12, 22, 48 and class II of ~~14, 27, 46, 35, 37 and 38~~. The visible portion of the maxillary sinuses is normally pneumatized. Regular course of the mandibular canal, predominantly lingual bilaterally, with emergence between 44 and 45 on the right, 34 and 35 on the left.

2026

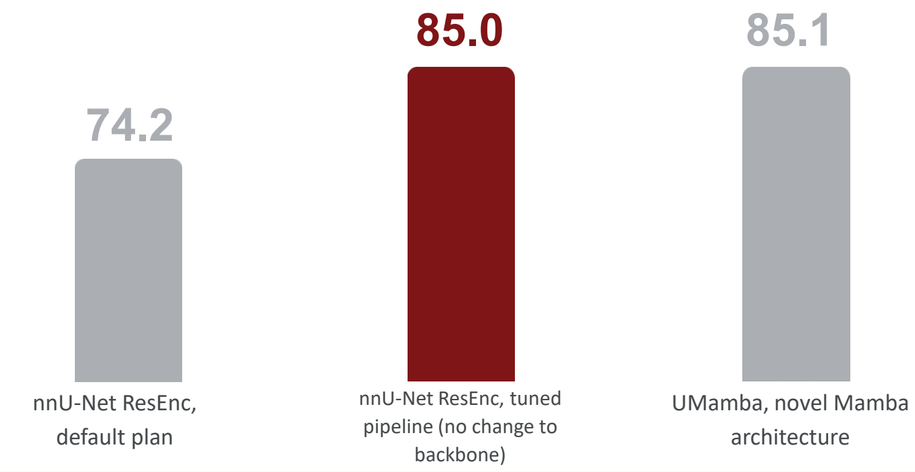
[1] Bolelli F., et al. "Segmenting the Inferior Alveolar Canal in CBCT Volumes: The ToothFairy Challenge" IEEE Transactions on Medical Imaging 2024.
[2] Bolelli F., et al. "Multi-structure segmentation in CBCT volumes: The ToothFairy2 challenge." Medical Image Analysis 2026.
[3] Lumetti L., et al. "ToothFairy3: Scaling CBCT Maxillofacial Segmentation to 77 Classes with U-Mamba2." MICCAI 2026.
[4] Lumetti L., et al. "Ontology-Grounded Structured Prediction for Dental CBCT Reporting." MICCAI 2026.

The nnU-Net Lesson

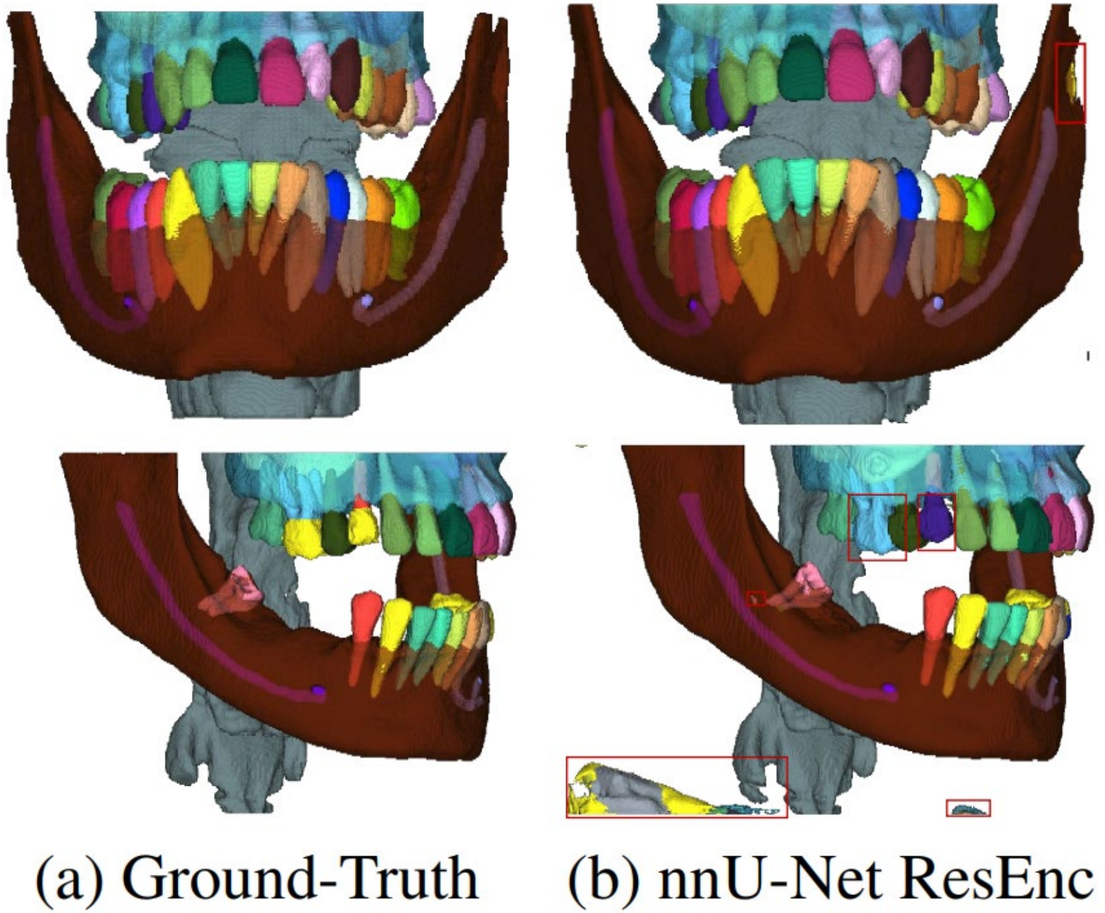
When methods are evaluated under the same protocol, strong nnU-Net-based pipelines remain extremely difficult to beat.

- This does not mean architectures do not matter.
- Once the benchmark is controlled, preprocessing, post-processing, training strategy and data quality can become at least as important as changing the backbone.

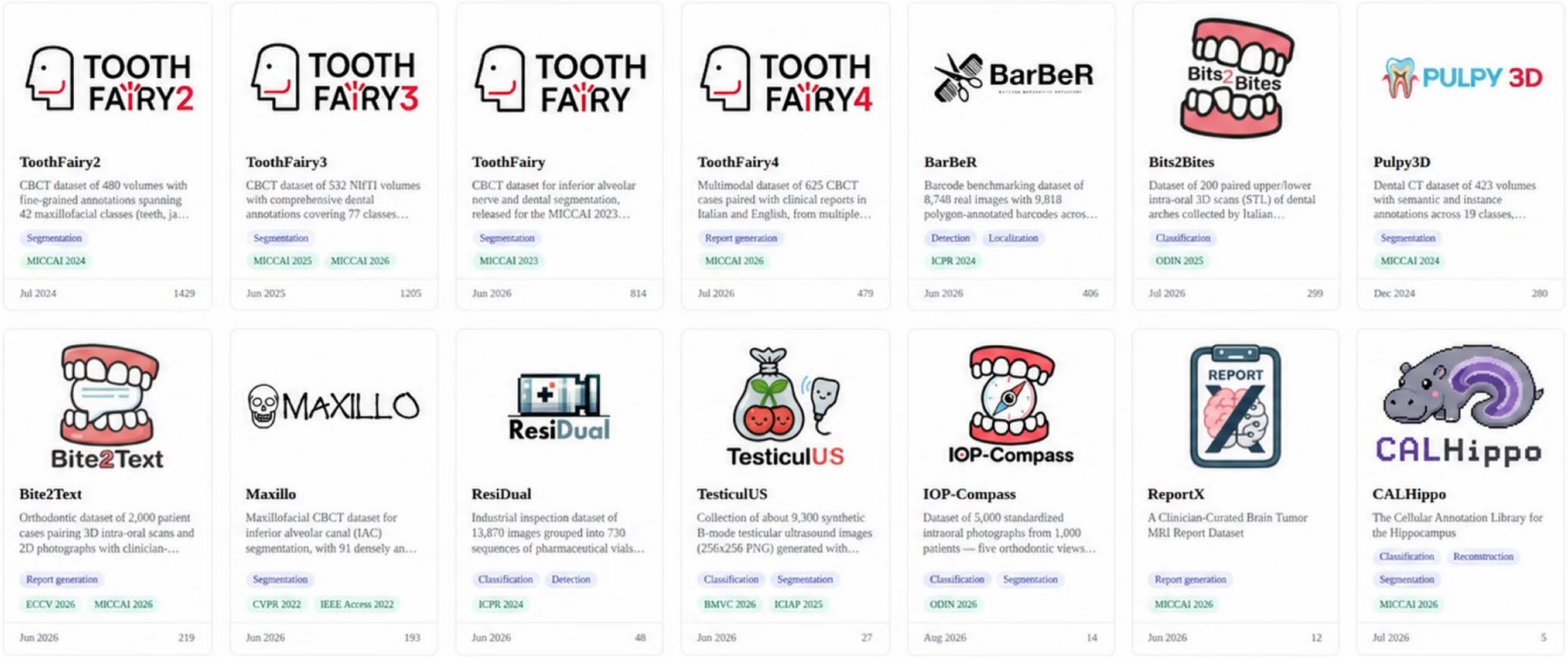
TOOTHFAIRY2 BENCHMARK — DICE SCORE (DSC, %)



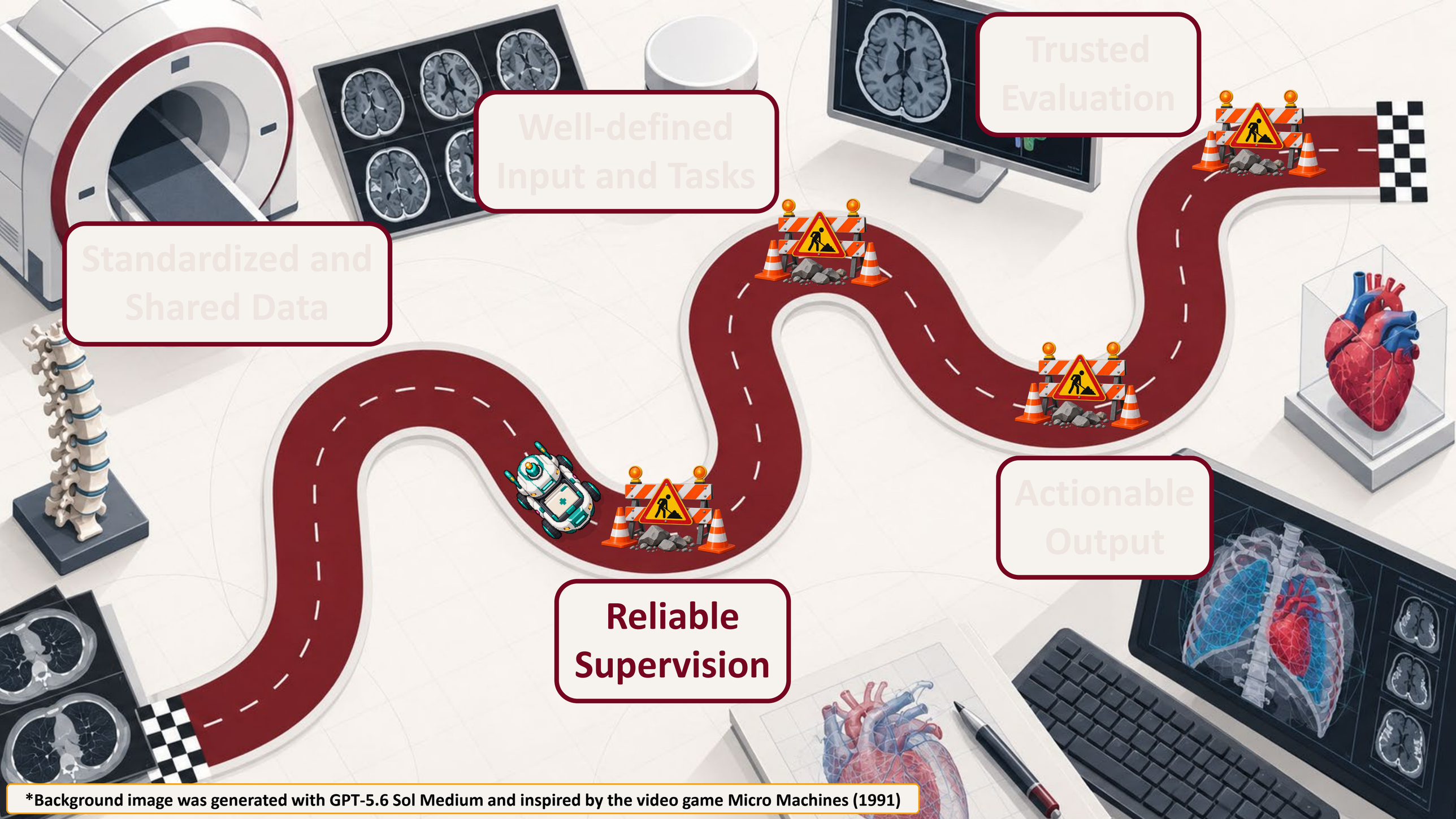
QUALITATIVE RESULT



[1] Bolelli, F., Marchesini K., et al. "Segmenting Maxillofacial Structures in CBCT Volumes". CVPR 2025.
[2] Isensee, F., Wald, T., Ulrich, C., et al. nnU-Net revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. MICCAI 2024.



Video not available in the PDF!



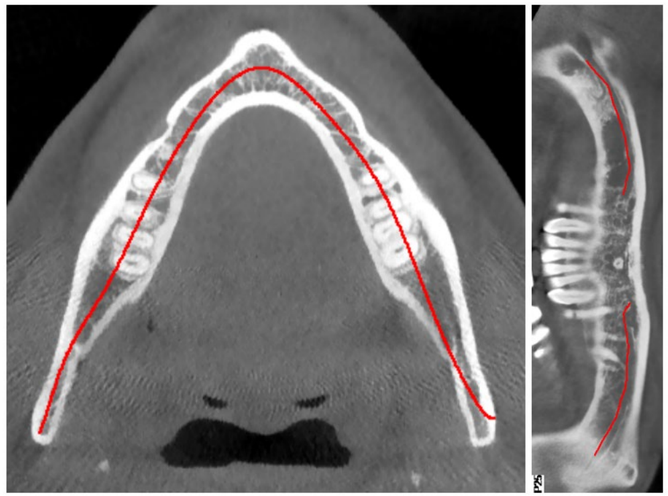
*Background image was generated with GPT-5.6 Sol Medium and inspired by the video game Micro Machines (1991)

Sparse Annotations Are the Data We Already Have

Dense 3D masks are the exception. Sparse labels are what clinics already produce.

01

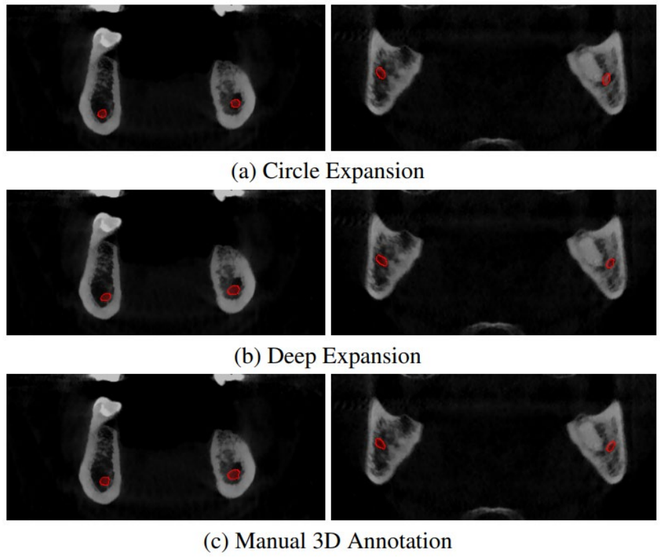
What the clinic gives you



SPARSE BY DEFAULT

02

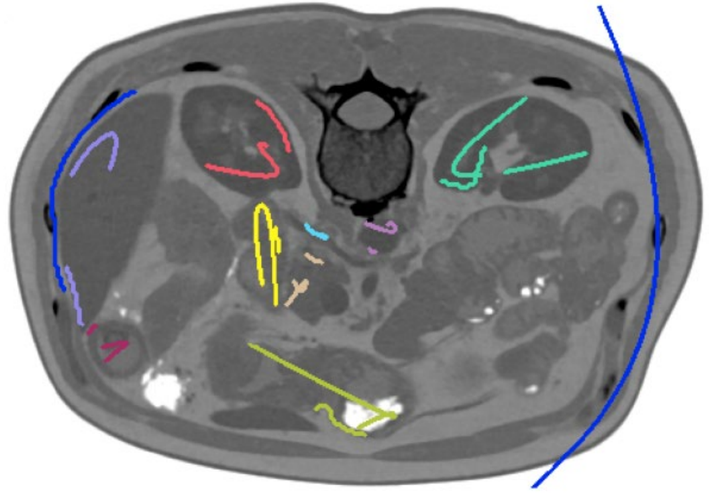
Propagate it



OUR ROUTE — CVPR 2022

03

Or ask for less



THE INTERACTIVE/SCRIBBLE ROUTE

[1] Cipriano M., et al. "Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation", CVPR 2022.
[2] Luo X., et al. "Scribble-Supervised Medical Image Segmentation via Dual-Branch Network and Dynamically Mixed Pseudo Labels Supervision", MICCAI 2022.
[3] Gotkowski, K. , et al. "Revisiting 3D Medical Scribble Supervision: Benchmarking Beyond Cardiac Segmentation", MICCAI 2025.

Reducing annotation time does not solve annotation consistency.

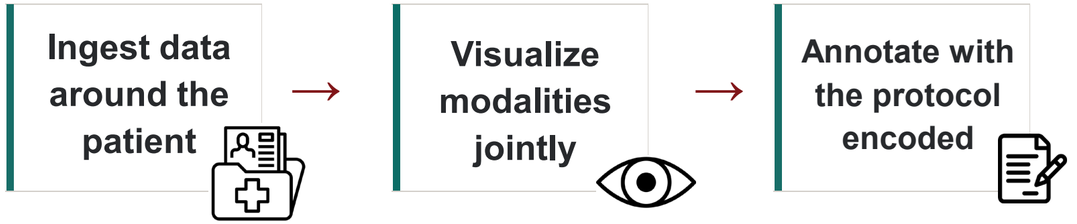
- We moved the protocol itself into the infrastructure.
- Data are organized around the patient. Data can be visualized jointly, and the annotation protocol is encoded in the platform.
- Clinicians work collaboratively, instead of producing independent datasets that we later try to reconcile.

Video not available in the PDF!

Yggdrasil

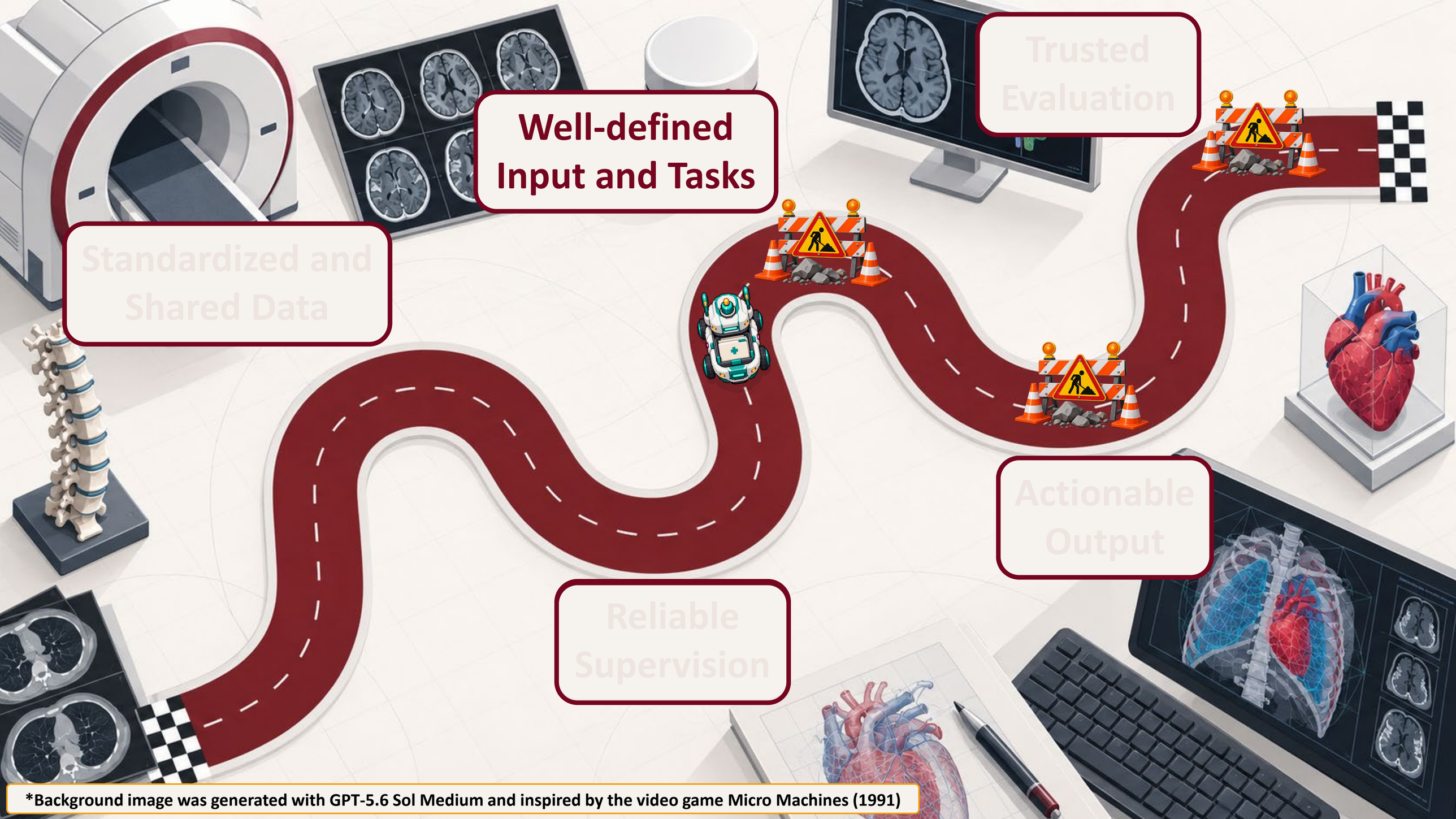
Multimodal clinical data and AI

AImageLab



Poster Session 2 — ExHall, Poster #325, 4:30 PM CEST

[1] Lumetti L., et al. "Do Multimodal LLMs Understand Intraoral Dental Data? Dataset, Platform, and Baselines", ECCV 2026.



*Background image was generated with GPT-5.6 Sol Medium and inspired by the video game Micro Machines (1991)

Missing Modalities Are the Normal Case

A model that requires every input to be present is a model that cannot be deployed.

WHY AN INPUT GOES MISSING

- Time in the acquisition slot
- Acquisition costs
- Patient condition
- Equipment available at that center

Accuracy should fall in proportion to the information lost — and the model should stay usable at every subset of inputs.

THE MRI EXAMPLE — BRATS

Brain tumor segmentation assumes four sequences — T1, T1c, T2, FLAIR. nnU-Net leads on the complete set, but the complete set is rarely what arrives. IM-Fuse degrades gracefully instead.

Mamba-based Fusion Model
IM-Fuse

[1] Pipoli V., Marchesini K., Saporita A., et al. "IM-Fuse: A Mamba-Based Fusion Block for Brain Tumor Segmentation with Incomplete Modalities", MICCAI 2025
[2] Carpentiero O., et al. "Brain Tumor Segmentation Under Missing Modality Scenarios: One Framework to Rule Them All", AAAI 2027 (Under Review)

The Task Itself Keeps Changing

A deployed model is not a finished artifact.

WHAT CHANGES AFTER DEPLOYMENT

New anatomical structures to cover

New scanners or centres appear

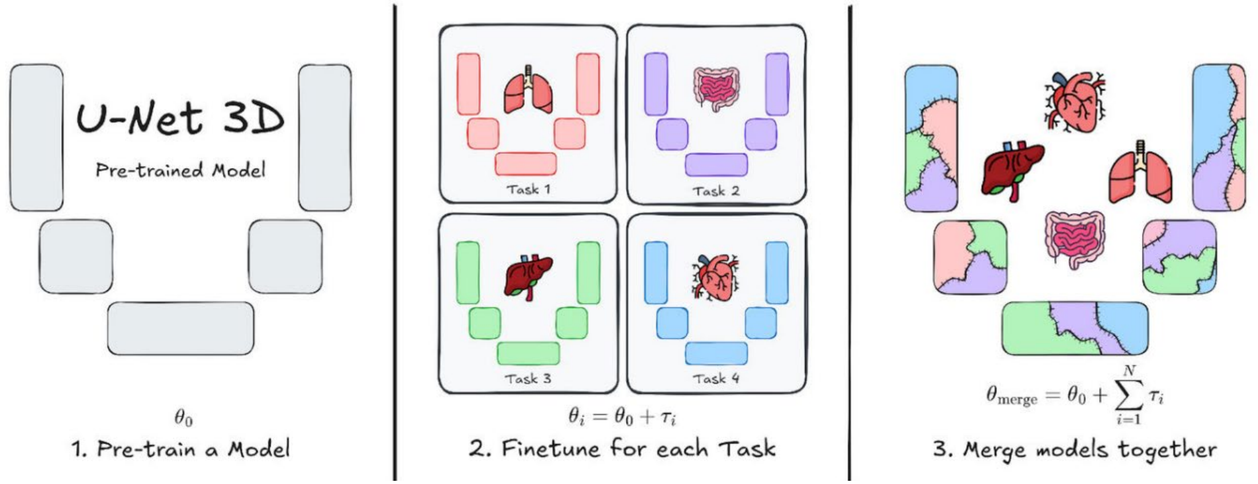
New clinical requirements emerge

The original training data are gone

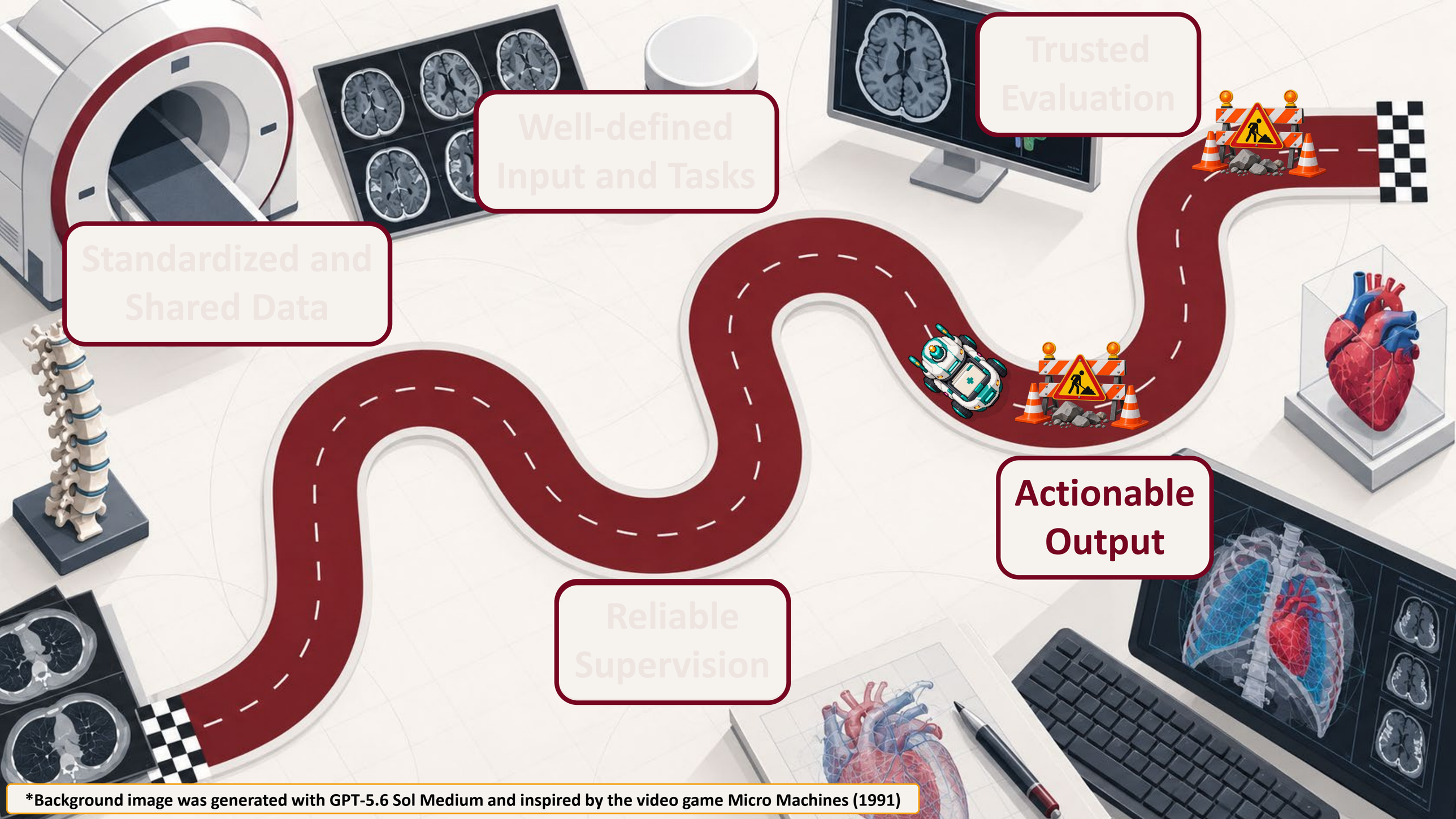
Adding a structure should not mean retraining everything from scratch.

THE CT/CBCT EXAMPLE — U-NET TRANSPLANT

Instead of one monolithic network, train specialists separately and merge them. Pre-training that encourages wide minima makes those task vectors combine cleanly — evaluated on ToothFairy2 and BTCV Abdomen.

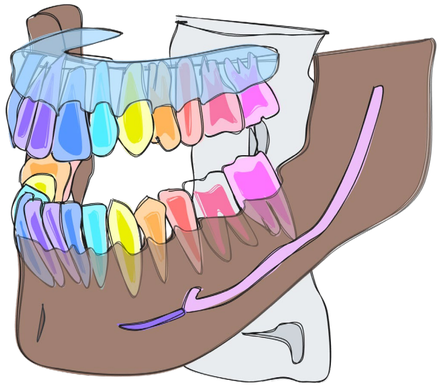


[1] Lumetti L., Capitani G., et. al "U-Net Transplant: The Role of Pre-training for Model Merging in 3D Medical Segmentation", MICCAI 2025.



*Background image was generated with GPT-5.6 Sol Medium and inspired by the video game Micro Machines (1991)

A patient is not a volume



SEGMENTED VOLUME

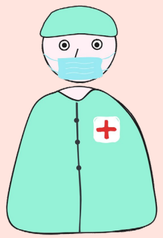
PATH A — RAW GEOMETRY

tooth 38

1.8 cm^3

Accurate, but not yet a clinical statement

→



Clinician's intervention

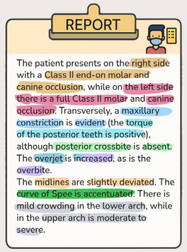
A human has to translate the numbers into findings

PATH B — REPORT GENERATION

Findings stated in clinical language

Relations, severity, laterality; what the volume implies

→



Directly actionable

The output now carries the clinical meaning

Segmentation describes anatomy. Reports communicate clinical meaning.

From Masks to Ontology-Grounded Reports

ToothFairy4 moves the benchmark from masks to structured findings.

WHAT WE RELEASED

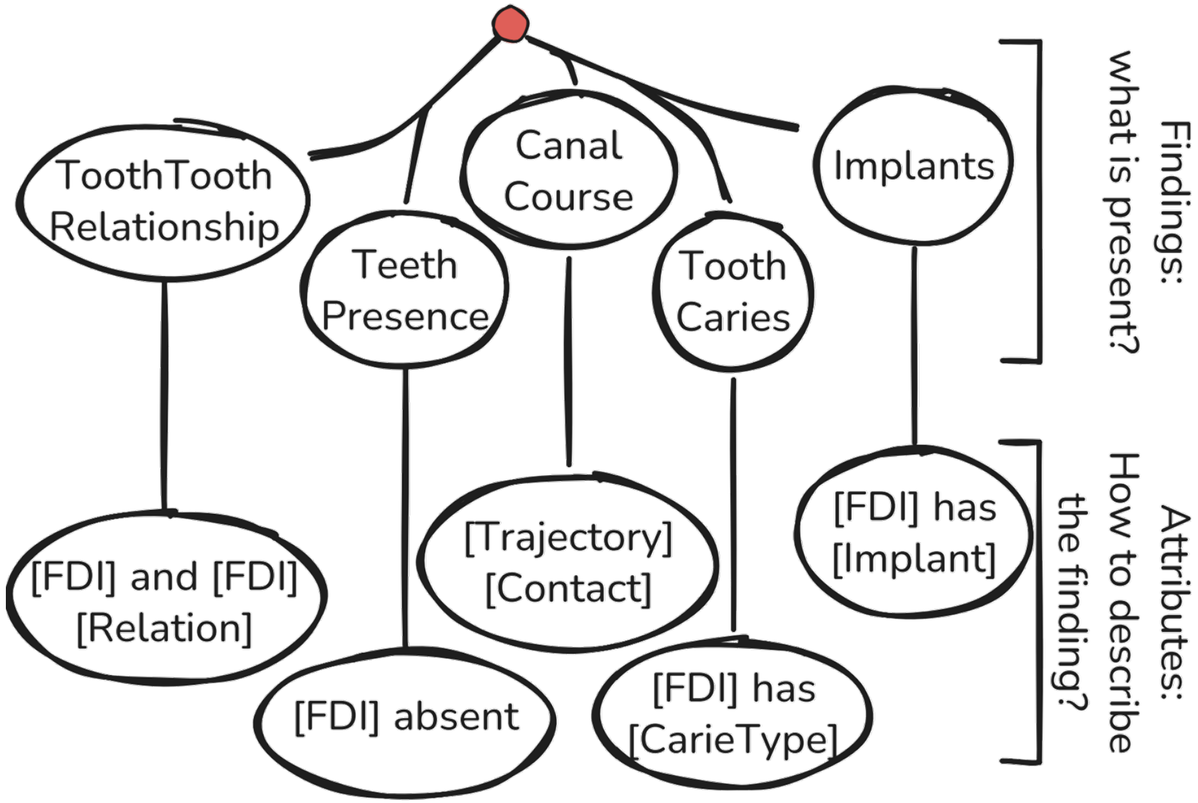
893 clinical reports for 529 public CBCT volumes

A clinician-designed OWL ontology, 13 finding types

Every report validated as an RDF/Turtle instance

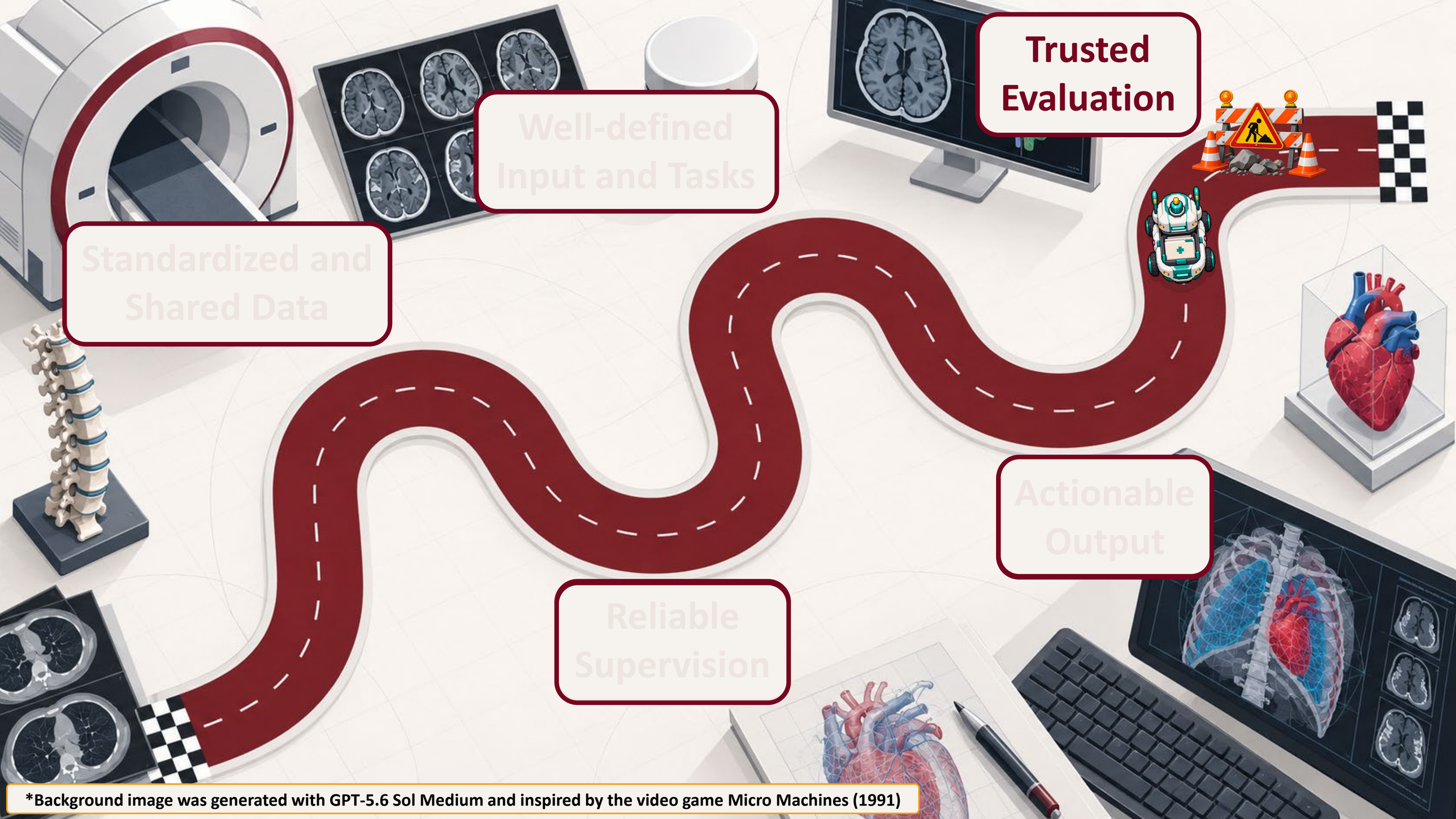
Six metrics that decouple the three stages

Reporting posed as structured prediction — not free text.



[1] Lumetti L., et al. "Ontology-Grounded Structured Prediction for Dental CBCT Reporting", MICCAI 2026.

[2] Saporita A., Pipoli V., et al. "PATHOCLASS-BRCA: Reframing PathologyReport Generation as Guideline-AlignedMulti-Task Classification", BMVC 2026



*Background image was generated with GPT-5.6 Sol Medium and inspired by the video game Micro Machines (1991)

Evaluation of Medical Findings

COMMON LEXICAL METRICS

BLEU n-gram precision

METEOR unigram alignment

ROUGE n-gram recall

All three score how a report is *phrased*. None of them check whether the finding is semantically **true**.

REF	Teeth	18,	28,	38,	48	are	**absent**
GENERATED	Teeth	18,	28,	38,	48	are	**endodontically treated**

6 of 7 tokens identical → high n-gram overlap

the one token that carries the diagnosis

0.68 **BLEU-4**

reads as a strong match

0.00 **FINDING AGREEMENT**

wrong diagnosis!

So maybe a more structured metric that checks actual findings is needed....

Evaluation of Medical Findings

RADFACT, SENTENCE LEVEL

Correctness
is each generated statement supported by the reference?

Completeness
does the report cover the reference findings?

LIMITATION

Both judgements are produced by an LLM — and inherit its failure modes.

● LLM JUDGE · SENTENCE-LEVEL ENTAILMENT

PROMPT

hypothesis "Absence of tooth 36"

reference
- edentulous 45
- edentulous 48
- edentulous 35
- edentulous 16
- edentulous 18
- restor. 46
- restor. 47
- endodontic 26
- post+core 17
- no entry: 36

JUDGE RESPONSE

label **entailment** → scored as correct

evidence "Edentulism of tooth 36"

never appears in the reference — invented

RADFACT STATUS
entailment
scored as correct

EVIDENCE IT CITES
Edentulism of tooth 36
never appears in the reference — invented

[1] Bannur, S., Bouzid, K., Castro, D. C., et al. (2024). MAIRA-2: Grounded Radiology Report Generation. arXiv:2406.04449. https://doi.org/10.48550/arXiv.2406.04449

Evaluation of Medical Findings

Video not available in the PDF!

AImageLab

Ranking Models Without Ground Truth

A hospital has ten candidate models, a new scanner and no annotations.

WHAT YOU DO NOT HAVE

- Labels for the target data
- The source training data
- Access to weights or features
- A published number you can trust

Published Dice on someone else's test set does not tell you which model to deploy.

RANK BY CONSISTENCY, NOT BY SCORE

Perturb the input and see whether the prediction survives. A model whose output stays invariant is the one that actually transferred Effective Invariance, extended from classification to semantic and instance segmentation.

UNSUPERVISED	SOURCE-FREE	BLACK-BOX

No labels, no source data, no weights — only the predictions.

Estimated rankings track true target-domain performance, in 2D and 3D.

Poster Session 4 — ExHall, Poster #391, 4:00 PM CEST

[1] Talks J., et al. "Unsupervised Source-Free Ranking of Biomedical Segmentation Models Under Distribution Shift", ECCV 2026.

PROGRESS, BUT NO PANACEA

Every direction helps — none of them is a panacea

01 Shared benchmarks solved comparison, not clinical evaluation

02 Human-in-the-loop can anchor clinicians to the model’s bias

03 Robustness is not clinic-ready, and multimodality demands more

04 Clinic-ready reports need blinded reads and better metrics

Our Research Group and Collaborations

AIMAGELAB — UNIVERSITY OF MODENA AND REGGIO EMILIA

GROUP


F. Bolelli


K. Marchesini


E. Candeloro


O. Carpentiero


N. Morelli


E. Vezzali


L. Borghi


G. Casari


G. Rosati


Y. Assefa


M. Lugli


L. Lumetti


C. Grana

CLINICAL PARTNERS

A. Anesi · M. Di Bartolomeo · A. Pellacani · P. Minafra · G. Pellacani · L. Bertoni · G. Bianchi · S. Puliatti · G. Besutti · R. Magistroni · F. Benuzzi · F. Lui · L. Lombardo · F. Cremonini

COLLABORATIONS

Radboudumc

Università degli Studi di Ferrara

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

杭州电子科技大学 HANGZHOU DIANZI UNIVERSITY

AImageLab

September 9, 2026

AI4M3D · ECCV 2026

ECCV 2026 | MALMÖ | SEPT 8–13

From Volumes to Clinical Decision

Five conditions to fill the gap between medical 3D vision and the clinic

Thank You!

Federico Bolelli

AImageLab — University of Modena and Reggio Emilia · federico.bolelli@unimore.it

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Papers, datasets, code